

ULRR

Modelling pain in population-based cohort studies: Biases, causal inference, and latent class methodologies

Item Type	Thesis
Authors	Ryan, Eva
Download date	2026-03-07 02:30:44
Item License	https://creativecommons.org/licenses/by-nc-sa/4.0/
Link to Item	https://doi.org/10.34961/researchrepository-ul.28219145

Modelling pain in population-based
cohort studies:
Biases, causal inference, and latent class
methodologies



Eva Ryan

Department of Mathematics & Statistics
University of Limerick

A thesis submitted for the award of Doctor of Philosophy
(Ph.D.)

Supervised by Dr. Helen Purtill and Prof. Ailish Hannigan

Submitted to the University of Limerick, October 2024

Declaration

This thesis is presented in fulfilment of the requirements for the award of Doctorate of Philosophy. I hereby declare that this thesis is my own work and that it has not been submitted for any other academic award. This work was completed under the supervision of Dr. Helen Purtill and Prof. Ailish Hannigan. Where the contributions of others were involved, every effort has been made to fully acknowledge and reference their work accordingly.

Signature:

Eva Ryan
October 2024

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Dr. Helen Purtill and Prof. Ailish Hannigan, for their continuous guidance, patience, and encouragement throughout my Ph.D. It was a privilege to have not one, but two, exceptional mentors, and your support made my overall Ph.D. experience very positive and enjoyable. I could not have asked for better supervisors!

I would also like to thank the collaborators I had the opportunity to work with during my Ph.D. A huge thank you to Prof. Zachary Zimmer and Dr. Hanna Grol-Prokopczyk, who graciously hosted my international research visits in Halifax, Nova Scotia and Buffalo, New York. A warm thank you also to our collaborators on the research trip project, Prof. Anna Zajacova and Dr. Christopher Dennison. I also wish to thank Dr. John Dziak and Dr. Bethany Bray in the University of Illinois at Chicago, and Dr. Peter May in Trinity College Dublin, whom I had the opportunity to work with earlier in my Ph.D. The generosity of each of my collaborators in sharing their time and knowledge greatly contributed to my development as a researcher, and showed me the huge benefits of working in an interdisciplinary team.

Next, thank you to all those in MACSI and the CRT, especially my fellow Ph.D. students and office mates, for making this experience such an enjoyable one. Meeting you has been the biggest highlight of the Ph.D. My sincere thanks is also due to Janet Clifford, Prof. James Gleeson, Peg Hanrahan, and Patsy Finn, for the incredible amount of work they put in behind the scenes for the CRT.

I would also like to say a big thank you to my friends outside of the CRT/MACSI, who helped to keep my stress levels low throughout the past four years. The moral support and laughs were much appreciated. Thank you also to my extended family - my grandparents, uncles, aunts, and cousins. A special mention to Nana and Dandan, who always showed an interest in what I was up to; and to my uncle Connie, who encouraged me to pursue a Ph.D. in the first place.

Finally, to my three biggest supporters – my parents Tina and Kieran, and my sister Leah – thank you for everything.

Publications

The following works form the basis of some chapters in this thesis.

Chapter 3 presents the following journal article:

Ryan, E., Hannigan, A., Grol-Prokopczyk, H., May, P., & Purtill, H. (2024). Sociodemographic disparities and potential biases in persistent pain estimates: Findings from 5 waves of the Irish Longitudinal Study on Ageing (TILDA). *European Journal of Pain* 28(5), 754–768. (Available online: <https://doi.org/10.1002/ejp.2215>).

Chapter 4 presents the following journal article:

Ryan, E., Grol-Prokopczyk, H., Dennison, C., Zajacova, A., & Zimmer, Z. (2024). Is the relationship between chronic pain and mortality causal? A propensity score analysis. *PAIN*[®] (Accepted and published online: <https://doi.org/10.1097/j.pain.0000000000003336>).

Chapter 5 is based in part on the following technical report:

Ryan, E., Dziak, J. J., Purtill, H., & Bray, B. C. (2023). Can a Normed Fit Index Assist with Model Selection in Latent Class Analysis with Large Samples? A Preliminary Investigation. (Technical report available on PsyArXiv: <https://doi.org/10.31234/osf.io/3qzvm>).

Abstract

Population-based cohort studies present unique opportunities to investigate research questions in areas such as population health and well-being. Addressing challenges in the statistical modelling of pain in older adults using cohort studies is the main goal of this thesis, motivated by two large cohort studies of ageing: the Irish Longitudinal Study on Ageing (TILDA) and the American Health and Retirement Study (HRS).

Firstly, the presence of pain-related attrition bias, mortality bias, and measurement bias due to differences in reporting styles (reporting heterogeneity) in TILDA is investigated. Evidence of mortality bias and reporting heterogeneity is found. Sex and socioeconomic disparities in pain previously reported in other countries are also observed in TILDA. Next, the causal effect of pain exposure on 20-year mortality in American older adults is estimated using HRS data. Propensity score methods are applied to adjust for measured confounding bias identified using a directed acyclic graph. Results suggest that pain likely causes a modest increase in mortality hazard, though the results are also compatible with no effect. Additionally, modifiable common causes of both pain and mortality are highlighted as potential targets for intervention.

Issues around the measurement of pain are also addressed. While pain is often modelled using a single measure such as pain intensity, it is desirable to identify a more holistic measure that incorporates multiple different aspects of pain experience. Latent class analysis (LCA) could be used for this task, however LCA model selection is particularly challenging for large datasets. The adaptation of fit indices from structural equation modelling for use with LCA is proposed to aid LCA model selection. The performance of the proposed indices is assessed using two simulation studies, and the indices show some potential when interpreted using an “elbow” rule. Finally, the proposed fit indices are applied to aid the development of an LCA model using various pain-related variables in the HRS dataset. Three distinct pain experience latent classes are identified, characterised by different patterns of pain impact and pain medication use. The latent classes are also found to differ across sociodemographic characteristics, with female sex and indicators of poorer socioeconomic background most common in the highest impact pain class.

In summary, this thesis addresses multiple challenges related to biases, causal inference, and latent class methodologies, developing approaches to strengthen the pain research evidence base. Findings are discussed within the context of current literature, and directions for future research are outlined.

Contents

	Page
1 Introduction	1
1.1 Population-based cohort studies and motivating datasets	1
1.2 Motivating health challenge	4
1.3 Potential biases and causal inference in population-based cohort studies	6
1.4 Latent variable modelling with large population-based cohort studies	10
1.5 Thesis structure	12
2 Background and Literature Review	15
2.1 Pain in older adults	15
2.2 Biases in cohort studies	20
2.3 Causal inference using observational data	25
2.4 Criticisms and limitations of causal inference using observational data	34
2.5 Latent variable modelling	37
2.6 Summary	41
3 Sociodemographic Disparities and Potential Biases in Persistent Pain Estimates: Findings from 5 Waves of The Irish Longitudinal Study on Ageing (TILDA)	45
3.1 Abstract	46

3.2	Introduction	47
3.3	Methods	49
3.4	Results	55
3.5	Discussion and conclusions	66
3.6	Acknowledgements	73
3.7	Ethics approval	73
3.8	Author contributions	73
3.9	Funding information	74
4	Is the Relationship Between Chronic Pain and Mortality	
	Causal? A Propensity Score Analysis	75
4.1	Abstract	76
4.2	Introduction	77
4.3	Methods	79
4.4	Results	91
4.5	Discussion	104
4.6	Acknowledgements	109
4.7	Appendix 1: Propensity score methods	111
4.8	Appendix 2: Sensitivity analyses	114
5	Normed Fit Indices for Latent Class Analysis with Large	
	Sample Sizes	121
5.1	Introduction	121
5.2	Latent class analysis (LCA)	123
	5.2.1 LCA model	123
	5.2.2 LCA model selection	127
5.3	Structural equation modelling (SEM)	131
	5.3.1 SEM model selection	131

5.4	Adapting SEM fit indices for LCA	134
5.5	Simulation studies	136
5.5.1	Simulation study 1	136
5.5.1.1	Design	137
5.5.1.2	Results	140
5.5.1.3	Discussion	154
5.5.2	Simulation study 2	157
5.5.2.1	Design	158
5.5.2.2	Results	161
5.5.2.3	Discussion	165
5.6	Conclusions	166
6	Pain Experience in American Older Adults: A Latent Class	
	Analysis	171
6.1	Introduction	171
6.2	Methods	174
6.3	Results	180
6.3.1	Model selection	180
6.3.2	Latent classes	183
6.4	Discussion	190
7	Conclusions and Future Research	195
7.1	Conclusions	195
7.2	Future research	200
A	Supplementary Material for Chapter 5	205
A.1	Simulation study 1: Results for unequal class sizes	205
A.2	Simulation study 2: NNFI results	214

B	Supplementary Material for Chapter 6	217
B.1	Summary of background characteristics for those with and without missing pain data	217
B.2	Bivariate residual summaries for the 2-class, 4-class, 5-class, and 6-class candidate LCA models	221
	Bibliography	223

1 Introduction

1.1 Population-based cohort studies and motivating datasets

Large-scale longitudinal population-based cohort studies (hereafter called cohort studies) involve the repeated collection of data from a cohort of individuals sampled from a defined population (Szklo, 1998). Cohort studies serve as rich sources of data on various aspects of life in the populations they are sampled from, presenting unique opportunities to investigate research questions in areas such as population health and well-being. Through effective sample recruitment and retention approaches, cohort studies aim to attract and retain a representative sample of the underlying population (Bonevski et al., 2014; Hartge, 2006). Representative samples achieve good external validity, and thus sample estimates derived from such studies are generalisable to the overall population (Szklo, 1998). Additionally, cohort studies are typically larger than clinical studies, often comprising of thousands or tens of thousands of participants. Such large sample sizes increase precision when estimating effects or associations between variables of interest. The collection of multiple waves of data on the same group of individuals is also a considerable advantage of cohort studies. Repeated measures enable the longitudinal modelling of health trajectories, providing invaluable insights into the development or progression of health outcomes for different population groups or subgroups over time (Smith and Dunn, 2022). Finally, since the recent introduction of

frameworks and statistical approaches to investigate causal relationships using observational data (Hernán and Robins, 2020) cohort studies have the potential to play an important role in expanding causal knowledge in health research (Hernán, 2018), particularly in situations where a clinical trial is not feasible due to financial constraints or ethical concerns.

However, accurate prediction and estimation of effects from large cohort studies is not without challenges. Firstly, selection bias may arise due to sample attrition if certain population groups are more likely to drop out or die during the study follow-up (Biele et al., 2019). Cohort studies typically handle such bias by using sample weights to account for attrition. However, it remains of interest to identify and understand the drivers of differential loss to follow-up or mortality in these studies. Measurement bias is another concern, particularly in health cohort studies where many measures are self-reported and thus subjective (Bago d’Uva et al., 2008). Confounding is also a potential source of bias, whereby failing to adjust for a common cause of the exposure and outcome under study can lead to spurious associations (Greenland et al., 1999). The aforementioned biases are particularly relevant when attempting to make causal inferences from observational data, a task which requires careful consideration and design to mitigate biases arising from selection, measurement and confounding (Hernán and Robins, 2020). Choosing appropriate statistical methods to model the data is another challenge of analysing cohort studies. It is often of interest to identify subgroups of individuals with similar health or sociodemographic backgrounds, however these subgroups are often not directly observable (i.e., are latent) and so must be estimated based on observed measures. Cohort study datasets often include many categorical variables, and so an appropriate method capable of handling this type of data must be chosen. Latent variable modelling methods such as latent class analysis and

latent growth curve modelling are often suitable methodologies (Collins and Lanza, 2010). However, latent variable model selection is a non-trivial task (Nylund et al., 2007), particularly when sample sizes are large.

The work in this thesis aims to address these challenges, motivated by two older adult cohort study datasets. The first is the Irish Longitudinal Study on Aging (TILDA; The Irish Longitudinal Study on Ageing, 2024) conducted by Trinity College Dublin. TILDA is a nationally representative sample of community-dwelling adults living in the Republic of Ireland, recruited using a two-step random sampling design. The purpose of the study is to collect information across a range of topics encompassing the health, social, and economic circumstances of its participants, providing a window into the experiences of those growing old in Ireland. The first wave of TILDA data collection was carried out between 2009 and 2010, with 8,171 adults aged 50 and over and 329 younger partners as respondents. Follow-up TILDA waves were conducted biennially, with the fifth wave of data collection occurring in 2018. A combination of personal interviews, self-completion questionnaires, and health assessments were used to collect data (Kenny et al., 2010).

The second motivating dataset for this thesis is the Health and Retirement Study (HRS; Health and Retirement Study, 2023) conducted by the Institute for Social Research at the University of Michigan. TILDA was modelled on the HRS, therefore the two studies are very similar in terms of design, content, and execution. The HRS began in 1992 with a cohort of American adults then aged 51-61 and their spouses of any age. The following year, the Assets and Health Dynamics Among the Oldest Old (AHEAD) study was launched to survey a cohort of Americans aged 70 and over at the time. In 1998, the HRS and AHEAD studies were merged along with two new cohorts, so that the HRS became a nationally representative sample of community-dwelling adults

aged over 50 in the contiguous USA. The sample was selected using a multi-stage area probability design with geographical stratification and oversampling of certain minority groups. New participants are recruited every six years to replenish the sample, with over 37,000 individuals having participated in the HRS by 2014 (Sonnegga et al., 2014).

Advances in data linkage have also increased the utility of these older adult cohort studies. For example, linkage with external mortality records allows researchers to model relationships between mortality and health-related variables in the study datasets, including the relationship between pain and mortality.

1.2 Motivating health challenge

Modelling pain in older adults using TILDA and HRS data is a central motivating challenge of this thesis. Defined as *“an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage”* by the International Association for the Study of Pain (Raja et al., 2020), pain can be characterised by a number of features including duration (acute or chronic), location (back, joint, widespread, etc.), and intensity (mild, moderate, severe, etc.) (Fillingim et al., 2016; Thienhaus and Cole, 2002). In the global network of older adult studies including TILDA and the HRS, pain is defined as being “often troubled by pain”. This definition is believed to capture persistent rather than transient pain (Grol-Prokopczyk, 2017), and will be the typical definition of pain used in this thesis. Persistent pain can be debilitating and have a negative impact on many aspects of life including quality of life, mental health, social connectivity, and physical functioning (Atkinson et al., 1991; Breivik et al., 2013; Cohen et al., 2021;

Dueñas et al., 2016; Hadi et al., 2019). Accurate estimates of pain prevalence, sociodemographic disparities in pain, and pain trajectories over time are crucial to guide policy makers and clinicians in reducing pain impact, particularly for ageing populations. Further, there is a scarcity of research exploring how pain *causally* affects health outcomes including mortality. A review of the literature found only two studies that explicitly investigate the causal effect of pain on mortality (Inoue et al., 2022; Smith et al., 2018), with most existing pain-mortality research discussing associations rather than causation (Smith et al., 2014). As investigations of the effects of pain exposure are confined to observational studies for ethical reasons, there is an opportunity to exploit recent developments in causal inference using this type of study design to further understanding of the pain-mortality relationship (Hernán and Robins, 2016). Additionally, using a single variable to capture pain experience (such as a measure of pain intensity or pain-related disability) can provide limited information. Some researchers have instead characterised pain by identifying latent subgroups of pain experience using latent class analysis (Dunn et al., 2006; O’Neill et al., 2020; Stynes et al., 2018), however selecting an optimal number of subgroups to identify is challenging. The statistical analyses of TILDA and HRS data in this thesis, outlined further in Section 1.5, are motivated by these challenges in pain research.

The following sections further introduce the challenges and opportunities presented by potential biases and causal inference (Section 1.3) and identifying subgroups and trajectories (Section 1.4) with cohort study data, using examples from pain research.

1.3 Potential biases and causal inference in population-based cohort studies

In longitudinal studies with multiple waves of follow-up, there is a risk of selection bias arising from attrition if certain population subgroups are more likely to be lost to follow-up or die over time (Biele et al., 2019; Metten et al., 2022). For example, one recent analysis of the Midlife in the United States cohort study found that high pain interference was associated with increased attrition risk, suggesting that those with greater pain interference become underrepresented in these samples over time (Liang, 2024). If ignored, selection bias can have serious implications for the accuracy and generalisability of results. Measurement error arising from the self-reported nature of many pain-related variables in cohort surveys is another potential source of bias. Systematic differences in reporting styles (reporting heterogeneity) have been observed for self-reported pain severity measures (Grol-Prokopczyk, 2017), and potential discrepancies in recall of pain intensity have also been explored (Dunn et al., 2010). Differential reporting styles and recall of experiences may reduce the reliability of self-reported measures, thus making it more difficult to accurately model trends and causal relationships using these measures. Finally, confounder bias is an important consideration when investigating relationships using observational data, particularly if a causal relationship is suspected. This type of bias is outlined more in the context of causal inference later in this section.

Throughout history, scientists have attempted to apply causal reasoning in their research to differentiate between association (where two variables are correlated but one does not influence the other) and causation (where one variable does influence or affect another variable). One landmark development

in the study of causation was the advent of randomised experiments, first formally introduced by Sir Ronald A. Fisher in the 1920s for agricultural studies (Fisher, 1925). In brief, randomisation balances the distribution of background characteristics in each exposure or treatment group, allowing any difference in outcome between groups to be attributed to the study intervention. The first modern clinical trial using randomisation is believed to have been conducted by epidemiologist Sir Austin Bradford Hill and colleagues in the Medical Research Council in 1948, to investigate the effect of streptomycin treatment in pulmonary tuberculosis (Medical Research Council, 1948). Bradford Hill later published nine “viewpoints” or criteria for consideration when assessing causality: (1) Strength of the association; (2) Consistency in the observation of the association; (3) Specificity of the association; (4) Temporality of the association; (5) A biological gradient (or “dose-response”) pattern to the association; (6) Plausibility of the suspected causal effect; (7) Coherence of the suspected causal effect with respect to existing knowledge; (8) Experimental evidence of causation; and (9) Analogy to an established causal relationship (Bradford Hill, 1965). The author emphasised that these “viewpoints” were not intended as a checklist of requirements for establishing evidence of causation. Rather, they are presented as relevant points to consider when evaluating if an observed association may be causal in nature. While the interpretation of each Bradford Hill criterion has evolved in line with advances in scientific knowledge, computing, and statistical methodologies since 1965 (Fedak et al., 2015), they continue to influence causal thinking and the development of causal inference approaches to this day. For example, many of the criteria can be mapped to modern directed acyclic graphs (Shimonovich et al., 2021). These graphs, which are further discussed later in this section, are an important tool used in the recently popularised potential outcomes (or

counterfactual) framework for causal research (Hernán and Robins, 2020).

Randomised controlled trials (RCTs) based on the randomised experiments formalised by Fisher (1925) are still widely considered the “gold standard” for investigating the effect of exposure versus non-exposure to a treatment of interest in health research. The ideal RCT conducted with a representative sample of the target population, effective randomisation to exposure groups, and complete adherence to study protocol over follow-up is the most stringent approach to establishing causal effects (Kendall, 2003). However, RCTs are not infallible. Poor study design and loss to follow-up can introduce biases that compromise the validity and generalisability of RCT results (Kendall, 2003). Additionally, some research questions cannot be addressed using an RCT due to safety or other concerns. For example, it would be unethical to expose trial participants to chronic pain to investigate its effects. In these cases, observational studies are the only option to investigate causal effects using data.

In recent decades, statistical methods and formal frameworks have been developed to elicit causal knowledge from observational studies under certain assumptions. The goal of these approaches is to emulate the target randomised trial that would ideally be carried out to investigate the causal relationship between treatment (or exposure) and outcome of interest (Hernán and Robins, 2016). In theory, this is achieved by identifying and adjusting for all common causes of the exposure and outcome under study to remove confounder bias, so that any difference in outcome estimated between exposure groups can then be solely attributed to exposure group. If all confounding is measurable and adjusted for, this process has the same effect as random assignment to exposure group in RCTs (Hernán and Robins, 2006).

In addition to confounder bias, researchers must also be aware of the selection and measurement biases mentioned earlier and mitigate them as much as possible to achieve accurate causal estimation from observational data. Directed acyclic graphs (DAGs) are useful tools for both clearly presenting the assumed causal structure underlying the analysis and identifying potential sources of confounder, selection, and measurement bias. Figure 1.1 is an example of a DAG to investigate the causal effect of pain on mortality. The exposure of interest (pain) is included in green, the outcome of interest (mortality) is included in blue, and the white boxes show confounders of the pain-mortality relationship that are adjusted for in the analysis. Directed arrows show the assumed flow of causality between relevant variables, while the absence of an arrow suggests no causal relationship. A more detailed discussion of the construction and interpretation of DAGs is given in Section 2.3.

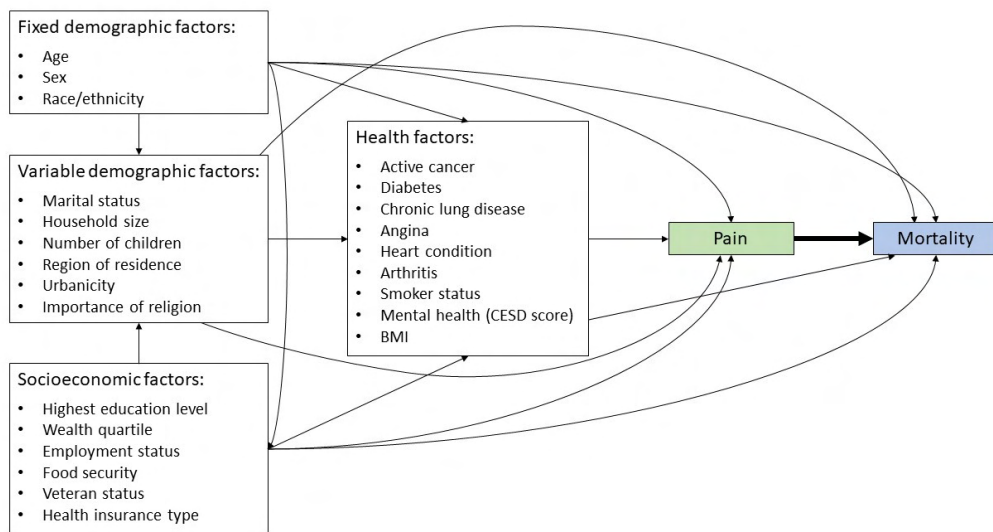


Figure 1.1: Example directed acyclic graph (DAG).

While this approach to causal inference using observational data opens new doors for causal investigation, potential pitfalls and criticisms must be

noted. Any inferences made are contingent on some unverifiable assumptions, including that the proposed causal structure is correct and that there is no unmeasured confounding. Additionally, experts caution that confounder selection should not be data driven, as adjusting for variables based on statistical correlations rather than consideration of their position in the causal structure may introduce bias (Van Zwieten et al., 2022). Rather, the selection must be informed by expert knowledge of the assumed underlying causal structure, ideally depicted using a DAG.

1.4 Latent variable modelling with large population-based cohort studies

As mentioned in Section 1.1, it is often of interest to identify subgroups of individuals with similar characteristics in cohort studies. Differences between subgroups can then be investigated to address important research questions, for example if pain outcomes differ between biopsychosocial subgroups (O’Neill et al., 2018). Subgroup membership is typically not directly observable, or latent, and so cluster analysis methods are applied to estimate the subgroups based on observed data. We note that previous publications use varied terminology including “profiles”, “phenotypes”, “patterns” and “typologies” to refer to these subgroups. For consistency, the term “subgroups” is used throughout this thesis.

Cohort studies often collect many categorical variables, and researchers must choose a clustering method suitable to the type of observed data available. Popular clustering methods designed for continuous data, such as k-means clustering and Gaussian mixture models, are not appropriate when using categorical indicator variables. Latent class analysis (LCA) is an alternative

model-based clustering method for identifying unobserved subgroups that is popular in the social, health and behavioural sciences. The cluster subgroups or “classes” are represented by the levels of a categorical latent variable that is estimated using observed categorical indicator variables (Collins and Lanza, 2010). For example, LCA has been applied to identify classes of lower back pain using pain variables from clinical assessment (Stynes et al., 2018). A key challenge for researchers applying LCA is the selection of an appropriate number of classes for the model to identify. Typically, a range of candidate models with different numbers of classes are fitted and then compared using model selection tools, such as information criteria and bootstrapped likelihood ratio testing (Collins and Lanza, 2010). However, for large sample sizes such as those seen in cohort studies, the traditional LCA model selection approaches can overestimate the number of classes required to adequately capture the underlying covariance structure in the data (Nylund et al., 2007). Favouring a model with more classes over a smaller, more practically useful model can result in interpretability or identifiability issues.

A range of longitudinal latent variable modelling methods have also been developed. Examples of applications in pain research include the use of ordinal latent growth curve modelling to estimate longitudinal sociodemographic disparities in pain severity (Grol-Prokopczyk, 2017), and the use of both longitudinal (or repeated measures) LCA and latent class growth analysis to model lower back pain trajectories (Kongsted et al., 2016). While there has been considerable latent trajectory modelling of lower back pain (Kongsted et al., 2016), further work investigating systematic differences in trajectories of other pain types using latent variable modelling could provide valuable insights for both clinicians and policy makers.

1.5 Thesis structure

This thesis explores various opportunities and challenges around the statistical and causal analysis of large cohort studies, with novel applications to pain research. In summary, the thesis structure is as follows. Chapter 2 consists of a literature review to contextualise the research in this thesis and provide background on the methodological approaches taken. The chapter begins with an overview of existing research around chronic pain in older adults (Section 2.1), focusing on sociodemographic disparities and the relationship between chronic pain and mortality in older adult populations. The remainder of Chapter 2 establishes the foundations of the statistical modelling concepts explored and utilised in this thesis. First, common biases that may arise when analysing longitudinal observational data are outlined (Section 2.2). Next, a formal framework for elucidating causal knowledge from observational data is introduced (Section 2.3). The final section of Chapter 2 details latent variable modelling approaches which have been used to identify pain subgroups and trajectories over time. An overview of the challenge of model selection and class enumeration in LCA models is provided (Section 2.5). Chapter 2 is a high-level overview of the relevant pain and statistical modelling literature. Additional detail on methodological and theoretical considerations specific to each project are included in the subsequent chapters.

The first goal of Chapter 3 is to model longitudinal sociodemographic disparities in pain in Irish older adults using 5 waves of TILDA data. Ordinal latent growth curve modelling is applied to estimate sociodemographic differences in pain trajectories. The second goal of this chapter is to explore potential biases arising due to attrition, mortality, or reporting heterogeneity, which may affect the accuracy of such pain estimates derived from longitudinal

observational data. The work in this chapter was published in the European Journal of Pain (Ryan et al., 2024).

Chapter 4 aims to investigate the causal effect of pain exposure on mortality over a 20-year period using HRS data. The assumed underlying causal structure is depicted using a DAG. Multiple imputation is used to handle missing data on the measured confounders. Measured confounding is then accounted for using propensity score matching or inverse probability weighting. Cox proportional hazard models are used to estimate the effect of pain on death as a hazard ratio. The work in this chapter was published in PAIN[®] (Ryan et al., 2024).

Chapter 5 addresses the challenge of selecting an appropriate number of classes in LCA when sample sizes are large, and traditional model selection methods may suggest an impractically large model that is uninterpretable or unidentifiable. The adaptation of normed and non-normed fit indices from structural equation modelling for use as LCA model selection tools is proposed. The behaviour and utility of the newly proposed indices are investigated using two simulation studies. Some of the work in this chapter is presented as a technical report available on PsyArXiv (Ryan et al., 2023).

Chapter 6 applies the LCA methodologies and new fit indices discussed in Chapter 5 to pain data from the HRS. The goal of Chapter 6 is to investigate if pain experience can be more holistically captured by pain latent classes identified using multiple pain variables, as an alternative to relying on a single measure to represent pain. This chapter also serves to test the behaviour of the LCA fit indices proposed in Chapter 5 when applied to a real world data example.

Finally, Chapter 7 provides an overall summary of the work in this thesis, along with some concluding remarks and directions for future research.

2 Background and Literature Review

This thesis aims to address key challenges and opportunities of longitudinal modelling in large observational datasets, with a focus on applications to pain research. Specifically, the developments in this work are motivated by improving the modelling of pain in older adults. A review of the literature pertaining to these key issues is conducted in this chapter. Firstly, some relevant research to date in the pain literature is presented. This is followed by an overview of different types of biases that can arise when analysing longitudinal observational data. Next, formal procedures for identifying and mitigating such biases when attempting to elicit causal knowledge from observational data are described. Criticisms and limitations of these procedures for conducting causal analyses are then discussed. Finally, some latent variable methodologies commonly used for modelling cohort study data cross-sectionally and longitudinally are introduced. In particular, issues arising around model fit for latent class analysis (LCA) are detailed.

2.1 Pain in older adults

The study of persistent pain in older adults is the main application explored in this thesis. Persistent pain is common in older adults, a group hereafter typically defined as those aged over 50 years. Some pain prevalence estimates

vary from 20-60% depending on location and pain definition (Dahlhamer, 2018; Zimmer et al., 2020). A body of research has shown that pain can have debilitating effects across many areas of life, including limiting functional ability (Breivik et al., 2013; Covinsky et al., 2009; Simmonds et al., 2012), reducing work performance (Blyth et al., 2003), increasing social isolation (Dueñas et al., 2016), negatively impacting mental health (Atkinson et al., 1991; Sheng et al., 2017), and potentially increasing mortality risk (Inoue et al., 2022; Smith et al., 2018). Pain as a health condition in older adults thus represents a considerable burden on the individual as well as on healthcare systems and society as a whole (Breivik et al., 2013). As life expectancy continues to increase globally, evidence-based planning will be required to manage the increasing numbers of older adults at risk of suffering from pain in the coming years (Gibson, 2007) and to minimise the impact of this pain. Consequently, understanding the sociodemographic drivers of persistent pain as well as the potential effects of this prevalent condition are pressing research concerns for the global ageing population.

In order to effectively tackle the burden of pain among older adults it is imperative to understand inequalities in pain, for example disproportionately high pain severity or pain-related disability in certain population groups. Sociodemographic disparities in chronic pain among adults of all ages have already been established, predominantly using cross-sectional analyses applied to cohort studies. Women and those from less advantaged socioeconomic backgrounds are typically observed to experience higher pain prevalence and greater pain severity on average compared to their counterparts (Mills et al., 2019; Van Hecke et al., 2013; Zajacova et al., 2021). Research specific to older adults on this topic is more sparse. One study of American older adults (aged 51+) also found similar sex and socioeconomic differences (as measured by

educational attainment and wealth) in pain experiences compared to studies of the general population (Grol-Prokopczyk, 2017). These disparities were modelled longitudinally using latent growth curve modelling, which primarily revealed differences in intercept (baseline pain) rather than slope (rate of change of pain over time). Further research investigating systematic differences in older adults' pain in other countries, particularly longitudinally, is needed to understand differential pain burdens in older adult populations internationally.

Additionally, further work is required to understand the relationship between pain and mortality in older adults. Similar to investigations of sociodemographic disparities, much of the existing research on pain and mortality consists of cross-sectional studies of adult cohorts. A systematic review of such studies conducted by Smith et al. (2014) found mixed evidence in the existing literature around the strength of the association between pain and mortality. Of the ten studies reviewed by Smith et al. (2014), six reported no significant association between pain and mortality after adjusting for relevant confounders (Andersson, 2009; Dreyer et al., 2010; Macfarlane et al., 2007; Smith et al., 2003; Torrance et al., 2010; Wolfe et al., 2011), while the other four studies reported a positive association between pain and mortality that remained after confounder adjustment (Macfarlane et al., 2001; McBeth et al., 2009; Nitter and Forseth, 2013; Sjøgren et al., 2009). Another apparently contradictory finding in cross-sectional pain research is the consistent increase in pain prevalence in cohort studies up to around 60 years of age, at which point pain prevalence appears to no longer increase with age (Domenichiello and Ramsden, 2019). Longitudinal modelling by Grol-Prokopczyk (2017) suggested that this apparent plateauing of pain was an artifact of cross-sectional analyses not accounting for disproportionately high mortality among those with more severe pain, rather than a genuine trend.

This finding highlights the importance of understanding potential attrition biases, and demonstrates a need for more longitudinal research to support understanding of pain.

There has been very little research explicitly examining the potentially causal relationship between pain and mortality. None of the studies discussed by Smith et al. (2014) apply a causal framework (Hernán and Robins, 2020), and use the term association rather than causation. A review of the literature found just two studies investigating the potential effect of pain on mortality in adults using an explicitly causal approach (Inoue et al., 2022; Smith et al., 2018). As it would be unethical to randomly expose participants to persistent pain in a clinical trial, both studies analyse observational data from cohort studies. Inoue et al. (2022) found that pain, mediated by opioid prescriptions, caused an increase in 3- and 5-year mortality in a group of American adults aged ≥ 20 . This study used the front-door formula, a useful and underused method for handling confounding under certain assumptions (Fulcher et al., 2020). The direct causal effect was not considered. Smith et al. (2018) analysed data from the English Longitudinal Study of Ageing (ELSA), a cohort study comparable to TILDA and the HRS. This study also found evidence of pain increasing mortality risk, mediated by a range of health, lifestyle, social, and psychological factors. The biopsychosocial model of chronic pain, first introduced by Engel (1977), provides a useful and holistic framework for modelling pain that incorporates its biological, social, and psychological dimensions. This model was adapted for older adults by Miaskowski et al. (2020), who highlighted factors including sex, age, race, socioeconomic status, depression and social support as key influences of pain in ageing populations. This adapted model could be used to inform the conceptualisation of future causal analyses involving pain exposure and to help

identify potential confounder variables. Such causal modelling considerations are discussed further in Section 2.3.

Pain research is also complicated by the challenge of effectively recording pain experience. Clinicians and researchers often rely on subjective self-reports, which raises concerns about the validity of such pain measures. For example, study participants' ability to accurately recall their pain experiences has been highlighted as one potential issue (Gendreau et al., 2003). However, reassuringly, some research has found that recall bias did not seriously impact the accuracy of retrospective pain self-reports. In one cohort of English adult primary care patients with back pain, good concurrence was found when comparing a retrospective report of pain severity for the preceding two weeks with daily pain diary entries for the same time period (Dunn et al., 2010). Strong agreement between recall of chronic pain for the preceding 12 months compared to repeated reports during the 12 month period was also observed for a cohort of Norwegian adults (Landmark et al., 2012). A related and possibly more serious pain reporting issue is reporting heterogeneity, whereby some population subgroups may have different reporting styles or perceptions compared to others (Bago d'Uva et al., 2008; Chan et al., 2011; Ziebarth, 2010). Such heterogeneity represents inconsistencies in the meaning of pain severity labels such as "mild pain", "moderate pain" etc. between groups, resulting in a type of measurement error that may bias pain analyses. Evidence, albeit tentative, has been found to support the hypothesis that there are systematic differences in pain reporting styles among American older adults (Grol-Prokopczyk, 2017). More research is required to further investigate the extent of pain reporting heterogeneity in older adult populations both in the USA and internationally. A third issue encountered when measuring pain is the difficulty in capturing pain experience using a single variable or a small number

of variables, e.g., pain severity or disability. Combinations of pain ratings have been found to provide the most accurate recall of pain experience for adults with low back pain (Dunn et al., 2010). Additionally, some researchers have used latent class methodologies to more holistically represent pain experiences as membership in unobserved pain subgroups (Dunn et al., 2006; O’Neill et al., 2020; Stynes et al., 2018).

This section provided a summary of key research and findings in the pain literature pertaining to sociodemographic disparities in pain, the potentially causal relationship between pain and mortality, and the challenges of effectively measuring pain. Numerous gaps in the literature and areas for further research using cohort studies were highlighted. The next section presents an overview of biases that may impact the accuracy of statistical or causal inferences made using cohort study data.

2.2 Biases in cohort studies

Cohort studies can be a valuable data resource for addressing research questions across the social, behavioural, and health sciences. However, care must be exercised at every stage of the data analysis process to identify and minimise the effects of potential biases that may compromise estimation accuracy (Ellenberg, 1994). Such biases can arise from a range of sources, including through measurement error during the data collection process (Brakenhoff et al., 2018) and through conditional selection of participants into the analytic sample due to attrition or mortality in longitudinal studies (Biele et al., 2019). Failure to mitigate such biases can lead to serious over- or underestimation of the estimand of interest. Examples include studies of the causal effect of statin use on cancer and on cardiovascular disease, where large

discrepancies have been found between estimates derived from observational data compared to randomised controlled trials (RCTs) (Danaei et al., 2012; Emilsson et al., 2018). As demonstrated by Dickerman et al. (2019) in the case of statin use and cancer, and by Danaei et al. (2013) for statin use and cardiovascular disease, careful study design can reduce bias and facilitate the estimation of causal effects from observational data which closely correspond to estimates derived from RCTs. This is achieved by emulating the design of an ideal or “target” trial when analysing observational data (Hernán and Robins, 2016), which is discussed further in Section 2.3. The remainder of this section discusses literature relevant to selection bias, measurement error bias and confounding.

Selection bias is a broad term covering a range of biases that arise from how individuals are selected into an analysis (Hernán and Robins, 2020). Sample attrition in longitudinal studies may result in bias if members of certain population subgroups are more likely than others to leave the study due to death or loss to follow-up (Biele et al., 2019). For example, differential attrition has been found to bias estimates of sociodemographic inequalities (Howe et al., 2013). Mortality bias arises if certain groups of participants have a higher mortality risk and are thus more likely to leave the sample due to death specifically (Grol-Prokopczyk, 2017). In some disciplines, the term “selection bias” is used to generally refer to this non-representativeness of certain groups in a sample (Smith, 2020). In other contexts, particularly when investigating causal effects, the term is used more specifically to refer to biases introduced by conditioning or stratifying on a specific variable. In the causal literature, selection biases are described structurally (Hernán et al., 2004) and can be broadly grouped into two categories (Lu et al., 2022). Firstly, collider stratification bias refers to selection bias arising due to conditioning

on a common effect of both the exposure and outcome of interest (Greenland, 2003). Berkson’s bias is a well-known case of collider stratification bias in statistics and epidemiology (Westreich, 2012). In his example, Berkson (1946) observed biased results when examining the relationship between an exposure and a disease after selecting study participants from attendees at a clinic, where clinic attendance was caused by both the exposure and disease. Secondly, selection bias can also be introduced into a causal analysis by restricting the analytic sample to just one or more levels of a mediator, or effect measure modifier, on the causal path between the exposure and outcome of interest (Lu et al., 2022). Such over-adjustment prevents the total causal effect of the exposure on the outcome being estimated consistently (Schisterman et al., 2009). The birth weight paradox is a common example of such bias, where stratifying on a mediator (birth weight) on the causal pathway between the exposure (maternal smoking) and outcome (infant mortality) of interest gives contradictory results: an apparently protective effect of maternal smoking on mortality risk in low birth weight infants (Hernández-Díaz et al., 2006). In the presence of selection bias, prevalence or causal effects estimated for the analytic sample will be different to the true prevalence or causal effects in the group eligible for selection. Thus, selection bias compromises the generalisability of the estimated values to the population of interest (Hernán et al., 2004; Nohr and Liew, 2018).

Measurement error bias can arise when there is a discrepancy between the value or concept of interest and the measured variable representing that concept in the collected data (Hernán and Cole, 2009), for example discrepancies due to data entry errors, inaccurate recordings, procedural mistakes, or differences in reporting styles (Van Smeden et al., 2020; Ziebarth, 2010). This type of bias is of particular concern in cohort studies, where

measures of health and well-being are often self-reported by participants rather than objectively measured. Specifically, such subjectively reported measures may be susceptible to reporting heterogeneity, or systematic between-subject differences in perceptions and reporting of experiences (Bago d’Uva et al., 2008). Consider a self-reported pain rating scale. The level of pain experienced when an individual reports “mild pain”, “moderate pain”, or “severe pain” may differ between sociodemographic groups if certain groups are on average more “stoical” than others in their self-reporting (Grol-Prokopczyk, 2017). For example, a more stoical individual might report “mild pain” for the same experienced level of pain that a less stoical individual would label “moderate pain”. Such measurement error can introduce bias, reduce precision, and mask data features, thus contributing to misleading or inaccurate results (Van Smeden et al., 2020). However, despite its ubiquity, measurement error bias is frequently ignored in epidemiological research (Brakenhoff et al., 2018; Jurek et al., 2006). A recent article by Van Smeden et al. (2020) highlights various misconceptions around measurement error, and provides suggestions for empirically investigating measurement error in epidemiological studies. Anchoring vignettes have also been proposed as a means to assess and account for reporting heterogeneity measurement bias specifically. Briefly, anchoring vignettes are hypothetical descriptions reflecting real-world experiences of individuals for the condition being studied (e.g., pain), and differences in how study participants rank or rate these vignettes provide insight into systematic differences in reporting styles (Molina, 2016). However, the utility of these vignettes depends on the assumptions of vignette equivalence and response consistency which may not always hold in practice (Grol-Prokopczyk et al., 2015). Identifying and addressing measurement error is thus an important but challenging task when investigating research questions using cohort studies.

Confounding is another key consideration when identifying sources of bias in analyses of cohort studies. In this thesis, the term confounding is used to describe a source of bias arising from a lack of comparability between the exposure groups under study when estimating causal effects, whereby the effects of the extraneous confounders are mixed with the causal effect of interest (Greenland et al., 1999). It is noted that other concepts are sometimes also included under the umbrella term of confounding, such as noncollapsibility of an association parameter for different levels of a covariate, and the inability to separate main effects and interactions for a particular study design (Greenland and Morgenstern, 2001). If ignored, confounding bias as defined in this thesis can lead to spurious associations or inflated causal effect estimates (Hernán and Robins, 2020). While confounding bias was traditionally handled by identifying and adjusting for potential confounders in a multivariable regression model, methods have also been developed specifically to account for confounding in causal analyses, such as those based on propensity scores (Austin, 2011). Statistical approaches to identify and adjust for confounding bias in causal analyses of observational data are discussed in more detail in Section 2.3.

This section provided an overview of biases that may arise when analysing observational data. In the next section, the literature of causal inference using observational data is introduced, establishing a framework for identifying and reducing the effects of biases, and proposed approaches to reduce or eliminate the threat these biases pose to accurate causal effect estimation are discussed. The assumptions underlying valid causal inference are also highlighted, and a broad introduction to the literature on eliciting causal knowledge from observational data is given.

2.3 Causal inference using observational data

As mentioned in Chapter 1, randomised controlled trials (RCTs) have long been considered the “gold standard” of causal research, as an ideal randomised experiment with complete adherence to protocol ensures that the treatment (or exposure) group and control group are exchangeable by design. Exchangeability refers to the state of both groups having similar distributions of relevant background characteristics and thus differing only in what treatment they receive (Hernán and Robins, 2020). In the presence of exchangeability, any differences in outcomes between the groups can be attributed to treatment assignment (Kendall, 2003), and it is this principle that allows for associations estimated from RCT data to be interpreted as causal effects. Exchangeability is explored in more detail below and the procedures that have been developed to conditionally achieve exchangeability in observational studies are discussed (Hernán and Robins, 2006). Two other identifiability conditions required for valid causal inference, consistency and positivity, are also briefly outlined (Hernán and Robins, 2020).

To explain **exchangeability**, the idea of a **counterfactual outcome** must first be introduced. Using the notation of Hernán and Robins (2020), we define a treatment (or exposure) variable A which for simplicity has two levels, $a = 1$ (receives treatment) and $a = 0$ (does not receive treatment). We also consider a dichotomous outcome variable Y with two levels, $y = 1$ (event e.g., death occurs) and $y = 0$ (event does not occur). We can then define two *potential outcomes* or *counterfactual outcomes*: $Y^{a=1}$, the outcome that would have been observed under treatment $a = 1$; and $Y^{a=0}$, the outcome that would have been observed under treatment $a = 0$. Note that as each individual will be assigned only one treatment value a , only one of the counterfactual values

$Y^{a=1}$ or $Y^{a=0}$ may actually be observed for each individual. Exchangeability means that the counterfactual risk of the outcome $Y^a = 1$ is the same in both the treated and untreated groups, i.e., $\Pr[Y^a = 1|A = 1] = \Pr[Y^a = 1|A = 0]$. In other words, the treatment groups are exchangeable if the counterfactual outcome Y^a is independent of actual treatment assignment A for all a . Thus, when exchangeability holds, the counterfactual risk under treatment for the group who actually received treatment is the same counterfactual risk as if the groups were swapped (or exchanged) and the untreated group instead received the treatment while the original treatment group were left untreated. Next, consider a confounder variable L . **Conditional exchangeability** holds if the treatment groups are exchangeable conditional on the confounder L , meaning $\Pr[Y^a = 1|A = 1, L] = \Pr[Y^a = 1|A = 0, L]$, or Y^a is independent of A given L . Causal inferences drawn from observational data are based on the assumption of conditional exchangeability i.e., that treatment groups are exchangeable after conditioning on measured confounders, and that there is no unmeasured confounding.

Measured confounders can be adjusted for using a number of approaches including regression adjustment, stratification, propensity score matching, and inverse probability weighting (Hernán and Robins, 2020). The latter two methods, which involve fitting a propensity score model, will be the main confounder adjustment methods considered in this thesis. Briefly, the propensity score is the estimated probability of receiving a particular treatment or exposure conditional on baseline characteristics (Austin, 2011). Thus, conditional on the propensity score, the distribution of background characteristics will be similar in both the treated and untreated groups. In propensity score matching, this balance is achieved by pairing treated and untreated individuals with similar propensity scores using a pre-specified

algorithm (Austin, 2014a). Inverse probability weighting involves calculating a weight for each participant using their propensity score, then applying these weights to create a synthetic sample in which treatment group is independent of baseline characteristics (Austin and Stuart, 2015). One benefit of using propensity score methods over traditional regression adjustment is that confounding is adjusted for separately using a propensity score model that is agnostic to the outcome, and thus the study design more closely reassembles an RCT (Amoah et al., 2020). Further, so called doubly robust methods have been developed which augment the standard propensity score approaches by also adjusting for confounding in the regression model for the outcome (Funk et al., 2011; Kreif et al., 2013). While propensity score matching or inverse probability weighting alone rely on the correct specification of the propensity score model to adequately adjust for confounding, when applying a doubly robust approach only one of either the propensity score model or outcome model need be correctly specified to obtain an unbiased causal estimate. The technical details of propensity score methods and doubly robust methods are discussed further in Chapter 4. Note that after applying propensity score methods, it is important to perform balance diagnostics to ensure confounding has actually been reduced and the distributions of the confounder variables in each exposure group are similar or balanced. Examining standardised differences between the means (or prevalences) of confounder variables in the exposed and unexposed groups is one common approach to assess balance (Austin, 2009a). Additionally, as mentioned previously, these adjustment methods can account for *measured* confounding only. If evidence of a causal effect is found, sensitivity analyses to examine the potential impact of *unmeasured* confounding on the estimated effect should be considered (D’Agostino McGowan, 2022).

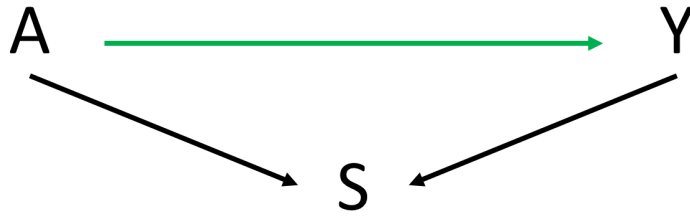
Identifying an adequate set of confounder variables to adjust for, with the aim of reducing confounding bias and achieving conditional exchangeability, is an important and often difficult task. As mentioned in Chapter 1, **directed acyclic graphs (DAGs)** are a useful visualisation tool to aid with confounder selection, as well as highlighting other potential sources of bias in the proposed analysis (Digitale et al., 2022). DAGs have been conceptualised as a mathematical language to visually represent both the subject-matter knowledge and statistical concepts underlying a hypothesised causal relationship (Pearl, 1995). They are composed of nodes (also called vertices) representing the different variables in the hypothesised causal structure, with arrows (also called directed edges) drawn between pairs of nodes to show the assumed direction of causal effects. The absence of an arrow between any pair of nodes implies no direct causal relationship between the two variables. A path is composed of a sequence of arrows connecting one variable to another, possibly through one or more other variables. The graphs are acyclic as there cannot be any cyclic paths leading from a variable back to that same variable, i.e., a variable cannot directly or indirectly cause itself. A causal path consists entirely of arrows pointing in the same direction, otherwise the path is non-causal. Considering an exposure variable A and outcome variable Y , a backdoor path is any non-causal path from A to Y that begins with an arrow pointing into A (Hernán and Robins, 2020).

Pearl (1993) proposed a “backdoor criterion” to select an admissible (or sufficient) confounder adjustment set L from a causal DAG based on two conditions. Firstly, variables in the set L cannot be descendant from A , meaning there cannot be a causal path from A to any variables in L . Secondly, adjusting for the set of variables L must “block” or “close” all backdoor paths from A to Y . The idea is that backdoor paths carry spurious associations

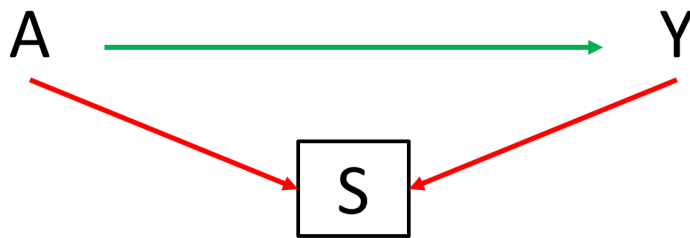
between A and Y , while causal paths pointing from A to Y carry causal associations (Pearl, 2010b). Thus, assuming the proposed causal structure is correct, blocking all backdoor paths removes confounding bias (achieving conditional exchangeability) and allows for the unbiased estimation of the causal effect of A on Y . The “flow” of association along a backdoor path can be blocked in a number of ways. One approach is to condition on a common cause of A and Y that falls on the backdoor path to close the path. Alternatively, if a collider variable falls on the backdoor path, this will block the flow of association. A collider is a common effect of two variables, such that the collider has two arrows pointing into it on the backdoor path. However, it is important to note that conditioning on a collider variable on a backdoor path will open the path, and should thus be avoided.

How to identify the selection and confounding biases discussed in Section 2.2 using DAGs and the backdoor criterion is demonstrated in the following simple examples. In these sample DAGs, the “flow” of the causal effect of the exposure A on the outcome Y along open direct and indirect paths is highlighted in green. Arrows comprising open backdoor paths between the exposure and outcome are highlighted in red. Conditioning or stratifying on a variable is depicted by drawing a box around that variable. C denotes a confounder of the exposure-outcome relationship; S denotes a common effect of both the exposure and outcome; and M denotes a mediator of the effect of the exposure on the outcome.

As mentioned previously, collider stratification selection bias is induced by conditioning on a common effect of both the exposure and outcome of interest, i.e., a factor which is downstream on causal paths from both the exposure and outcome (Hernán et al., 2004). Consider the DAG in Figure 2.1(a) which has no open backdoor paths between exposure A and outcome Y . As depicted



(a) Without conditioning on S , there are no open backdoor paths between the exposure A and outcome Y .



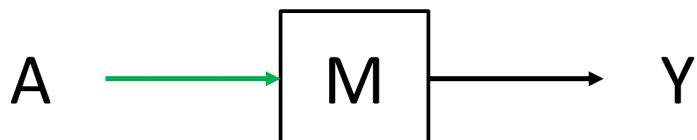
(b) Stratifying on the collider S opens the backdoor path $A \rightarrow S \leftarrow Y$.

Figure 2.1: Example of collider stratification bias represented using DAGs.

in Figure 2.1(b), conditioning on the collider S opens a backdoor path $Y \rightarrow S \leftarrow A$, leading to collider stratification bias. Similarly, while the DAG in Figure 2.2(a) does not depict any biases, conditioning on the mediator M as in Figure 2.2(b) “blocks” the path from A to Y . This selection prevents the accurate estimation of the effect of A on Y by introducing selection bias. In the final example in Figure 2.3(a), bias arises due to failure to adjust for a confounder variable rather than unnecessarily adjusting for a collider or mediator variable as in the previous two examples. There is an open backdoor path $Y \leftarrow L \rightarrow A$, which can be closed by conditioning on the confounder L as depicted in the DAG in Figure 2.3(b). This removes confounder bias and, assuming the causal structure represented in the DAG is correct, allows for the unbiased estimation of the causal effect of A on Y .



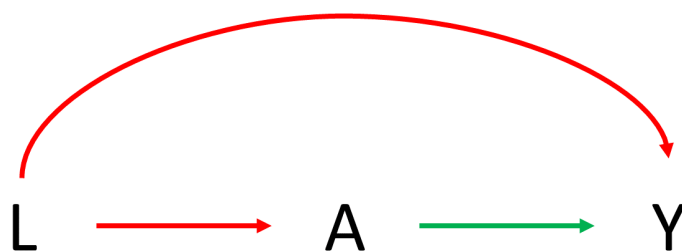
(a) The effect of the exposure A on the outcome Y is mediated through M .



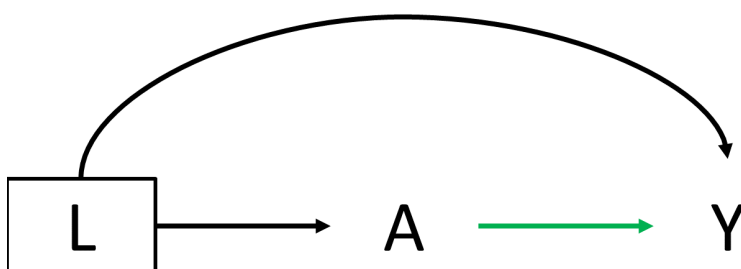
(b) Conditioning on the mediator M “blocks” the causal path from A to Y .

Figure 2.2: Example of selection bias introduced by conditioning on a mediator of the exposure-outcome relationship represented using DAGs.

The accuracy of any causal inference made using the counterfactual framework depends on the validity of the assumed underlying causal structure and the adequate identification and control of relevant biases. Thus, creating a DAG to visualise this structure is an important step in causal study design and requires careful consideration. However, determining what variables to include and their position in the DAG is a non-trivial task, and must be based in substantive expert knowledge rather than statistical criteria (Hernán et al., 2002). Automated approaches to identify statistical associations between other variables and the exposure or outcome of interest, such as statistical testing using p-values, generally cannot distinguish between confounders and mediators of the total causal effect (VanderWeele, 2019). As discussed previously, while adjusting for confounders reduces bias, adjusting for mediators introduces bias when estimating the total causal effect (Van Zwieten et al., 2022). Although data-driven approaches to confounder selection can



(a) The exposure A and outcome Y share the common cause L . Failing to condition on L leaves the backdoor path $Y \leftarrow L \rightarrow A$ open, and so estimation of the causal effect of A on Y will be subject to confounder bias.



(b) Conditioning on L removes confounder bias by blocking the backdoor path $Y \leftarrow L \rightarrow A$.

Figure 2.3: Example of confounder bias represented using DAGs.

be helpful in some high-dimensional cases, they should not supersede expert knowledge (VanderWeele, 2019). Regarding the presentation of DAGs, a recent review by Tennant et al. (2021) explored common practices in applied health research articles, and based on their findings the authors provided best practice guidelines to increase transparency and utility when creating DAGs. Their recommendations included clearly stating the focal causal relationship and estimand of interest as well as the DAG-implied confounder adjustment set for the estimand; including all relevant variables even if they are not directly measured; arranging the DAG so that all causal arrows flow in the same direction; justifying alternative adjustment sets and reporting their results separately; and generally assuming that a causal arrow exists between any

two variables, as omission of an arrow is a much stronger assumption than inclusion.

While central to the potential outcomes framework of counterfactual reasoning, exchangeability achieved by controlling for measured confounders is not the only condition required for valid causal inference using observational data. **Consistency** is also needed for the identification of causal effects (Hernán and Robins, 2020). The consistency assumption states that a participant’s potential outcome under the treatment (or exposure level) they actually receive should be the same as the observed outcome for that participant (Cole and Frangakis, 2009; Pearl, 2010a; VanderWeele, 2009a). Using the earlier counterfactual notation, this assumption can be written as $Y^a = Y$ for any individual who receives treatment $A = a$. The consistency criteria encapsulates the idea that there should be no different versions of the treatment $A = a$ such that the potential outcome for Y is different depending on the version of $A = a$ received (Rehkopf et al., 2016). Ambiguity in the definitions of treatment or exposure levels should thus be avoided to preserve consistency. Additionally, for consistency to hold there should also be no interference, meaning that an individual’s counterfactual outcome does not depend on the exposure value for any other individual (Hernán, 2012; Hernán and Robins, 2020).

Positivity, also called the experimental treatment assignment assumption (Mortimer et al., 2005), is another necessary assumption for causal inference using observational data. Positivity holds when every possible combination of values on the observed confounders is experienced in both the exposed and unexposed groups in the population of interest (Westreich and Cole, 2010). Using the notation of treatment A and confounder set L required for exchangeability, the positivity condition can be expressed as the requirement that $Pr[A = a|L = l] > 0$ for all treatment values $a \in A$ and all values l

for which $Pr[L = l] \neq 0$. Thus, positivity requires that conditional on the confounder set L , the probability of receiving each treatment is non-zero, or positive (Hernán and Robins, 2020). Having both exposed and unexposed participants within each confounder strata is important to enable comparison. There are two types of positivity violations: structural (or theoretical) violations, where it is theoretically impossible for certain population subgroups to receive a particular treatment a ; and practical (or random) violations, where receiving every treatment level is theoretically possible within each subgroup, but some treatment levels are not received in certain subgroups in the study due to chance (Zhu et al., 2021). The latter violation can be thought of as a small sample size problem (Petersen et al., 2012) which can be ameliorated by increasing sample size.

This section provided a broad overview of the literature on counterfactual reasoning and established frameworks for eliciting causal knowledge from observational data, such as the data collected by large cohort studies. The next section describes some criticisms and limitations of these approaches to causal investigation as presented in the literature.

2.4 Criticisms and limitations of causal inference using observational data

While DAGs and the counterfactual framework are powerful tools for investigating causal questions, they are not without issues and criticism. For example, concerns have been expressed that causal inference in epidemiology may become too narrowly focused on this one conceptual framework, to the exclusion of other valuable views of causal reasoning (Krieger and Davey Smith, 2016; Vandenbroucke et al., 2016). These critics have also suggested the

counterfactual framework is too restrictive, as counterfactuals may only be defined in terms of feasibly modifiable interventions (e.g., medication dosage) and thus meaningful counterfactuals cannot be defined for potential causes that cannot be conceptualised as humanly modifiable (e.g., race) (Vandenbroucke et al., 2016). Robins and Weissman (2016) countered this criticism by highlighting that when the definition of a causal effect is ambiguous enough to prevent agreement about its meaning, the only way to reduce this vagueness is to more precisely define the hypothetical intervention. Robins and Weissman (2016) suggested that this criticism is directed at the counterfactual framework specifically merely due to the transparency with which the framework presents this issue, which applies to all types of causal analyses.

Consistency violations, whereby the treatment (or exposure) is ill-defined such that different versions of the treatment lead to different outcomes (Rehkopf et al., 2016), can also compromise the validity of counterfactual causal analyses. Concerns about the ability to clarify or unambiguously define certain types of exposure has been highlighted as a challenge to the adaptation of causal inference methods in social epidemiology specifically (Glymour and Rudolph, 2016). Causal experts have clarified that while there should be no meaningful vagueness in specifying the exposure/treatment of interest, absolute precision is not required (Hernán, 2016) and common misconceptions should not prevent the integration of counterfactual causal inference in social epidemiology (Galea and Hernán, 2020).

Another limitation of applying the potential outcomes framework is the unverifiable assumption of no unmeasured confounding (Hernán and Robins, 2020). While complete elimination of confounder bias cannot be guaranteed in any causal analysis, careful consideration and presentation of the study design and the assumed underlying causal structure improves transparency and

facilitates open appraisal of the validity of the causal inferences made (Hernán, 2016; Tennant et al., 2021). Sensitivity analyses have also been developed as useful tools for evaluating the robustness of the causal estimate to violations of the unmeasured confounding assumption, by determining the degree of unmeasured confounding required to nullify the causal estimate (Cinelli and Hazlett, 2020; D’Agostino McGowan, 2022; VanderWeele and Ding, 2017).

As discussed by Krieger and Davey Smith (2016), no one approach can unequivocally establish the existence of a causal effect, and causal triangulation is required to build a compelling case based on a broad range of evidence. Triangulation involves using a range of different study designs and methodological approaches applied in different populations to investigate the causal relationship of interest. As each analysis will have its own statistical assumptions and be subject to different sources and directions of bias, triangulation is a means to evaluate the robustness of causal estimates and provides a stronger basis for causal inference (Hammerton and Munafò, 2021).

Despite the limitations and challenges of making causal inferences based on observational data using a counterfactual approach, the potential outcomes framework provides an important opportunity to explore causal relationships which would be otherwise infeasible or even impossible to study e.g. in the study of the effects of chronic pain experience as an exposure on outcomes such as mortality. Applying a counterfactual approach to cohort study data represents an opportunity to nonetheless investigate this and other important research questions.

The final section of this chapter provides an overview of the literature of another key methodological approach used in this thesis; latent variable modelling.

2.5 Latent variable modelling

Latent variable models are useful tools for modelling underlying phenomena which cannot be directly measured. Such latent variables can represent important but abstract hypothetical constructs of interest to researchers, such as worker efficiency or intelligence (Bollen, 2002). Typically, the latent variable is assumed to be causally related to measurable factors, called manifest items or indicators, and it is by collecting data on these indicator variables that a model is fit to estimate the structure of the underlying latent variable of interest (Bollen and Hoyle, 2012; Collins and Lanza, 2010). A variety of methods have been developed for modelling continuous, discrete, and categorical latent variables both cross-sectionally and longitudinally (Muthén, 2002). As the focus of this thesis is cohort studies which often collect substantial amounts of categorical data, an overview of the literature of latent variable modelling approaches that can be estimated using categorical indicator variables will be presented in this section. Note that modelling methods for continuous data which fall under the umbrella of structural equation modelling, such as factor analysis, are discussed briefly in Chapter 5.

Latent class analysis (LCA) has been developed as an approach to estimate the structure of a categorical latent variable at one time point (Collins and Lanza, 2010). Since the first book providing a comprehensive introduction of LCA by Lazarsfeld and Henry (1968), this methodology has grown in popularity and been applied across a range of fields including pain research (Dunn et al., 2006; O’Neill et al., 2018; O’Neill et al., 2020; Stynes et al., 2018). The categories of the latent variable identified using LCA, called “classes”, characterise subgroups in the population of interest which cannot be observed or recorded directly. These classes may represent important

conceptual subgroups which are central to pain research questions but not directly measurable. Examples include identifying biopsychosocial risk factor subgroups for pain development in older adults (O’Neill et al., 2018), and subgroups of clinical relevance such as characterising individuals with low back-related leg pain (Stynes et al., 2018). LCA is thus a valuable tool which facilitates the investigation of important research questions involving abstract concepts that would be unidentifiable if relying on direct observation. Briefly, the LCA model is characterised by two sets of parameters: latent class membership probabilities, which estimate the proportion of the population belonging to each latent class; and class-specific item response probabilities, which estimate the probability of individuals within each latent class selecting a particular response on each of the indicator variables (Collins and Lanza, 2010). Covariates (e.g., sex, age) may also be included in the LCA model to examine the association between these covariates and latent class membership (Collins and Lanza, 2010). A technical overview of LCA modelling and assumptions is provided in Chapter 5.

Assessing model fit is a primary concern in all statistical modelling to ensure the selection of a final model which adequately represents the covariance structure underlying the observed data (Kadane and Lazar, 2004). Evaluating model fit is a particularly important consideration in LCA, as it encapsulates the task of selecting an appropriate number of classes to identify when estimating the latent class variable. Also called “class enumeration”, the goal is to select the smallest number of classes such that the covariance structure in the data is parsimoniously captured by the chosen LCA model (Collins and Lanza, 2010). LCA can be conceptualised as a type of model-based clustering (Gormley et al., 2023), where the classes of the categorical latent variable represent clusters in the population. Class enumeration is thus

directly comparable to the task of manually selecting an appropriate number of clusters when performing other types of cluster analysis, such as parametric Gaussian model-based clustering (Fraley and Raftery, 2002; Grün, 2019) or unparametric k-means clustering (James et al., 2013). For LCA, typically the first step in selecting an appropriate number of clusters/classes is to fit multiple candidate models with different numbers of classes. The quality of model fit for each candidate model is then compared to determine the number of classes that gives the most parsimonious fit to the data (Nylund-Gibson and Choi, 2018). Similar to other parametric clustering methods (Gormley et al., 2023), LCA model fit is often evaluated using information criteria such as the Akaike Information Criterion (AIC; Akaike, 1973) and Bayesian Information Criterion (BIC; Schwarz, 1978), or goodness-of-fit hypothesis testing such as the bootstrap likelihood ratio test (BLRT; McLachlan and Peel, 2000). However, these common approaches can fail to select the most parsimonious model in practice when sample sizes are large. For example, the AIC has been observed to overestimate the number of classes required when conducting LCA with larger samples (Nylund et al., 2007). Similarly, when analysing large samples the p-values calculated for the BLRT can suggest a need to add more classes, even when discrepancies between the fitted and hypothesised model are small and not practically significant. This is a specific case of a commonly observed and well known phenomena in statistical significance testing, whereby very small violations can give “statistically significant” results in hypothesis tests when applied to large samples (Greenland et al., 2016; Lin et al., 2013). The increase in number of parameters to be estimated associated with additional classes can cause identifiability issues, while a large amount of smaller classes can make it difficult to meaningfully interpret the fitted LCA model (Collins and Lanza, 2010). Thus, parsimonious LCA class enumeration

for large sample sizes is an important ongoing area of study.

Latent class methodologies have also been extended from using cross-sectional data to analysing longitudinal data. One such approach is repeated measures latent class analysis (RMLCA), where a latent class model is fitted using responses recorded for the indicator variables at multiple time points. The estimated classes then represent different patterns of changes over discrete time points. RMLCA is thought to perform best when there are a small number of indicators recorded at three or more time points (Collins and Lanza, 2010). Latent transition analysis (LTA) is another form of longitudinal latent variable modelling for categorical data. Rather than modelling responses to the indicator variables at all time points simultaneously (as for RMLCA), LTA typically estimates a baseline LCA model using data from the first time point, then characterises changes in response patterns over time as transitions in latent class membership between adjacent discrete time points (Lanza et al., 2003; Nylund-Gibson et al., 2023). RMLCA and LTA are both useful approaches when the research goal is to investigate unobservable behaviour profiles and changes in these patterns over time, when the changes are assumed to be discontinuous (Nylund-Gibson et al., 2023).

Latent growth curve modelling (LGCM) is an alternative approach to using latent variable methodology with longitudinal data (Duncan et al., 2013). Considered a special form of structural equation modelling (Preacher et al., 2008), the goal of LGCM is to model continuous latent growth trajectories based on observed measurements at multiple discrete time points. LGCM has the benefit of estimating change in growth trajectories both between and within individuals over time (Duncan and Duncan, 2009). The basic latent growth model is characterised by an intercept factor and a slope factor, which are allowed to co-vary. The intercept factor gives the mean and variance of the

intercepts for each individual growth curve, while the slope factor details the mean and variance of the change in individual growth trajectories over time (Duncan et al., 2013). Multivariate extensions to the basic model facilitate investigations of how the growth trajectories vary within levels of relevant predictor variables (Curran et al., 2010). While originally developed to model continuous response variables, the LGCM approach has also been adapted for ordinal categorical response variables (Masyn et al., 2014; Mehta et al., 2004). For example, ordinal LGCM has been used to model trajectories of pain and sociodemographic disparities in pain (Grol-Prokopczyk, 2017) based on ordinal response variables. Ordinal LGCM is conceptualised as a model of a continuous underlying latent growth trajectory which is coarsely measured by the observed ordinal variable (Duncan et al., 2013). Threshold values are estimated by the model to mark the partitioning of the continuous latent variable into the categories of the ordinal variable (Masyn et al., 2014). Unlike RMLCA and LTA which model change at discrete time points only, ordinal LGCM assumes continuous change which is modeled by the slope factor. Further details of the ordinal latent growth curve model are provided in Chapter 3 of this thesis.

2.6 Summary

The purpose of this chapter was to provide relevant context and background information for the work conducted in this thesis. The chapter began by summarising research to date in the literature pertaining to the key pain topics addressed: sociodemographic disparities in pain in older adults, the potentially causal nature of the relationship between pain and mortality, and challenges in measuring pain for statistical analysis. Next, an overview of selection, measurement, and confounder biases that may impact the accuracy of pain

estimates derived from cohort studies was provided. This was followed by a summary of key concepts and tools relevant to the counterfactual approach to causal inference using observational data. Finally, the chapter concluded by introducing the latent class methodologies applied in this thesis, along with the particular challenge of selecting an optimal number of classes for an LCA model with large sample sizes.

This literature review highlighted numerous knowledge gaps requiring more research. There is a need for further investigation of sociodemographic disparities in pain, especially longitudinally and in cohorts outside of the United States. Also, careful consideration must be given to potential selection, measurement, and confounder biases that may impact pain estimates. Efforts to identify, quantify, and mitigate these biases are crucial, particularly when interested in investigating causal effects. This literature review also highlighted a lack of research investigating the potential causal effect of pain on mortality, with nearly all existing studies of the pain-mortality relationship estimating associations rather than applying appropriate methods to estimate causal effects. Finally, the challenge of effectively measuring pain was discussed. It was found that LCA represents a potentially useful tool for identifying more holistic measures of pain experience, however the selection of an optimal LCA model being complicated by large sample sizes presents another issue.

Chapter 3 addresses a number of these challenges by investigating the presence of pain-related attrition, mortality, and measurement bias in the TILDA dataset, while also contributing new knowledge of longitudinal sociodemographic disparities in pain in Irish older adults to the pain literature. Chapter 4 addresses the lack of research explicitly investigating the causal relationship between pain and mortality by conducting a formal causal analysis using HRS data, and in the process provides a novel template for conducting

causal analyses involving a pain exposure using propensity score methods. Next, Chapter 5 tackles the challenge of LCA class enumeration for large sample sizes by proposing and testing the adaptation of fit indices from structural equation modelling for use in LCA. These fit indices are then used in Chapter 6 to address the challenge of pain measurement by applying LCA to HRS data to identify a holistic pain measure that encompasses different aspects of the pain experience. Finally, Chapter 7 provides conclusions and discusses findings within the context of the literature, as well as outlining directions for future research.

3 Sociodemographic Disparities and Potential Biases in Persistent Pain Estimates: Findings from 5 Waves of The Irish Longitudinal Study on Ageing (TILDA)

The work in this chapter has been published in the *European Journal of Pain*:

Ryan, E., Hannigan, A., Grol-Prokopczyk, H., May, P., & Purtill, H. (2024). Sociodemographic disparities and potential biases in persistent pain estimates: Findings from 5 waves of the Irish Longitudinal Study on Ageing (TILDA). *European Journal of Pain* 28(5), 754-768 (Available online: <https://doi.org/10.1002/ejp.2215>).

The content of the article is reprinted verbatim in this section. Details of all author contributions are included in the article text. The author of this thesis (ER) was responsible for designing and carrying out the analysis, interpreting the results, writing the first draft of the manuscript, revising the draft, and

managing the submission process.

3.1 Abstract

Background: Pain is a prevalent, debilitating condition among older adults. Much evidence on this topic comes from cohort studies, which may be affected by attrition and measurement bias. Little is known about the impact of these biases on pain estimates for European older adults. Additionally, there is a lack of longitudinal research on pain and sociodemographic disparities in Irish older adults.

Methods: We analysed data from 8,171 participants (aged ≥ 50 at baseline) across five waves of The Irish Longitudinal Study on Ageing. Longitudinal pain severity and sociodemographic disparities in pain were explored visually and using a latent growth curve model. Using multivariate logistic regression, we examined bias due to attrition at later waves associated with reported pain at Wave 1. Measurement biases due to reporting heterogeneity were assessed by investigating associations between sociodemographic factors and pain-related disability for given pain levels.

Results: Wave 1 severe pain was associated with increased odds of attrition due to death by Wave 5 (AOR: 1.63, 95% CI: 1.20, 2.19). Not having private health insurance was associated with increased odds of pain-related disability at Wave 1, controlling for pain severity (AOR: 1.37, 95% CI: 1.15, 1.64). These results suggested mortality bias and reporting heterogeneity measurement bias respectively. Sex, education level, and private health insurance status disparities in pain were observed longitudinally.

Conclusions: Mortality bias and reporting heterogeneity measurement bias must be accounted for to improve older adult pain estimates. There is a need

for policymakers to address sociodemographic disparities in older adult pain levels.

3.2 Introduction

Pain is estimated to affect 30-60% of adults aged ≥ 50 years in Europe (Zimmer et al., 2020). Negative effects of pain include isolation, disability, and reduced quality of life (Breivik et al., 2013; Cohen et al., 2021). Internationally, studies of “older adults” (defined here as those aged 50 and older) typically find that the burden of pain is greater for certain sociodemographic subgroups, such as women and socioeconomically disadvantaged groups (e.g., those with lower education attainment) (Cimas et al., 2018; Stewart Williams et al., 2015).

Much evidence about pain comes from population cohort studies. However, these can be subject to attrition biases (Biele et al., 2019; Metten et al., 2022) if participants with certain characteristics are more likely to be lost to follow-up or die, which may affect estimates of pain prevalence, severity, or disparities. For example, large-scale volunteer databases of ageing are less likely to retain less healthy, less socioeconomically advantaged participants (Brayne and Moffitt, 2022). Mortality bias arises if excess attrition of certain sociodemographic subgroups occurs specifically due to death.

Grol-Prokopczyk (2017) explored the potential impact of bias due to mortality or non-response when estimating sociodemographic disparities in pain for older American adults, using multi-wave data. While no evidence of association between pain and non-response was found, pain severity was strongly predictive of mortality, suggesting that pain estimates may be subject to mortality bias. A similar association between severe pain and increased mortality risk was found in a Scottish older adult cohort (Torrance et al.,

2010).

Measurement bias arises from systematic differences in self-reporting styles (reporting heterogeneity). There is a growing literature examining sociodemographic differences in reporting styles for subjectively rated health conditions, including pain (Bago d’Uva et al., 2008; Chan et al., 2011; Ziebarth, 2010). Failure to recognize differences in reporting styles could impact the validity of relative rankings and comparisons between groups (Menec et al., 2007). Estimating the presence and direction of such self-reporting bias is thus an important task, sometimes undertaken by comparing self-reports to more objectively recorded measures of condition (Jürges, 2007; Spitzer and Weber, 2019).

Limited research has investigated the impact of biases due to attrition and specifically mortality (Lacey et al., 2013; Muszyńska-Spielauer and Spielauer, 2022) and reporting heterogeneity (Spitzer and Weber, 2019) on health status estimates among *European* adults. To our knowledge, the impact of these biases on estimates of *sociodemographic disparities in pain* has not been explored in any European cohorts. Research on the demography of pain in Ireland in particular is sparse, consisting of a small number of cross-sectional and longitudinal analyses examining associations between older adult pain and factors including health and healthcare utilisation (Kennedy et al., 2017; O’Neill et al., 2018; O’Neill et al., 2020; Raftery et al., 2011).

Establishing accurate, unbiased estimates of pain prevalence and sociodemographic disparities is crucial to inform public policies targeting this potentially pervasive and debilitating condition. This study uses five waves of the Irish Longitudinal Study on Ageing (TILDA) to contribute to this goal by (1) examining previously unexplored longitudinal sociodemographic disparities in pain among older Irish adults and (2) investigating how attrition

bias, mortality bias, and reporting heterogeneity measurement bias may affect the accuracy of such pain estimates.

3.3 Methods

Population and participants

This study is a secondary analysis of five consecutive waves of TILDA. TILDA is a nationally representative cohort study of the health, social, and economic conditions of community-dwelling older adults in the Republic of Ireland. A multi-stage sampling design was used to select the baseline (Wave 1) sample. The first step involved grouping all residential addresses in the Republic of Ireland into 3,155 townland clusters, using the Irish Geodirectory as a sampling frame. These clusters were stratified by socioeconomic group and geography, and a representative sample of 640 clusters was selected. A probability sample of 40 addresses was then drawn from each cluster and contact was made to recruit household members aged 50 and over. A response rate of 62% was achieved at the household level, with 8,171 community dwelling older adults and 329 of their younger partners participating in the study at Wave 1 made available for analysis. Wave 1 commenced in 2010 and subsequent waves of data collection occurred every two years. At each wave, participants were invited to complete a computer-assisted personal interview, a self-completion questionnaire, and a health assessment. Health assessments were carried out in one of two TILDA health centres or in the participant's home. The full study design is detailed elsewhere (Kenny et al., 2010).

Individuals who were aged 50 or over at Wave 1 of TILDA were included in the longitudinal analysis. Younger partners were excluded. Of the 8,171 participants in our Wave 1 sample, 6,993 (85.6%) returned for Wave 2; 6,246

(76.4%) for Wave 3; 5,571 (68.2%) for Wave 4; and 4,872 (59.6%) for Wave 5. This closed cohort design ensured that longitudinal data across the five waves would be present for all participants, except in the case of non-response or death. We defined attrition as participants leaving the cohort between waves, either due to death or other loss to follow-up.

Measures

Pain:

At each TILDA wave, participants were asked, “Are you often troubled with pain?” (yes, no). We label this pain phenotype “persistent troubling pain” (“pain” for brevity). While a duration of pain is not specified, previous research suggests this question is unlikely to capture acute or transient pain. One study found respondents were more than twice as likely to report “any pain in the last 30 days” as to report being “often troubled by pain” (Banks et al., 2009). Those who answered yes to this initial pain question were then asked, “How bad is the pain most of the time?” (mild, moderate, severe). Responses to both pain questions were combined to make a 4-category “pain status” variable for each wave. For some parts of the analysis, these pain status categories were converted to a numerical pain score using the following codes: 0 = no pain, 1 = mild pain, 2 = moderate pain, 3 = severe pain, as in previous studies (Dunn et al., 2006; Grol-Prokopczyk, 2017), and scores averaged across groups. Those who reported being often troubled by pain were also asked “Does the pain make it difficult for you to do your usual activities such as household chores or work?” (yes, no). This question was used as an indicator of experiencing pain-related disability or not. These pain questions have been used previously as measures of experiencing pain, pain severity, and pain disability respectively in older Irish adults (O’Neill et al., 2020) and in

older adult populations worldwide (Bell et al., 2022; Mohanty et al., 2022).

This study focused on *non-cancer* pain only. Responses to pain variables were set to missing in cases where reported pain was likely due to cancer or cancer treatment. This rule affected between 0.1% and 1.2% of the sample at each wave.

Sociodemographic factors:

Demographic variables were age category (50-59, 60-69, 70-79, 80+), sex (male, female), highest level of education (none or primary, secondary, tertiary), and whether individuals were covered by a private health insurance policy (yes, no) at baseline. Education level and private health insurance status were used as proxies for socioeconomic status (SES); while TILDA does collect data on income and assets, these are answered only by a subset of the sample; education and insurance are answered by all participants and therefore by using them we minimise missing data issues. Irish healthcare services are financed both publically and privately, with the primary benefit of private health insurance being reduced wait times for elective hospital treatments (Turner et al., 2020). 47% of the Irish population had private health insurance at the end 2021 (Health Insurance Authority, 2023).

Attrition by Wave 5:

Attrition by Wave 5 was defined as someone not participating in Wave 5 for any reason, including death. As part of sample maintenance efforts the TILDA team attempted to contact and invite baseline participants for interview at each follow-up wave even if they had missed a previous wave, unless the participant had been confirmed deceased or requested to withdraw from the study (Donoghue et al., 2017). If someone missed an intermediate

wave/waves after Wave 1 but had returned to the study by Wave 5 they were included as present in the Wave 5 sample. The number of Wave 5 participants who had missed at least one prior wave was small ($n = 278$, 5.7% of those present at Wave 5).

Mortality data:

Mortality data included survival status (confirmed deceased or not confirmed deceased) and year of death if confirmed deceased. To obtain this data, the TILDA team performed linkage between General Register Office death records and individual level survey data. The full linkage process is detailed elsewhere (Ward et al., 2020). The linkage identified 741 deaths in our sample between the end of Wave 1 and March 2018 (Wave 5 data collection commenced in January 2018). By comparing year of death to year of data collection for each wave after Wave 1, a survival status variable was created for each follow-up Wave (2-5), indicating whether a participant was still alive at that wave. When a participant died between waves, their survival status was set to alive at the prior wave and deceased at all following waves. For example, a participant who died in 2013 was coded as alive at Wave 2 (collected in 2012) and deceased from Wave 3 (collected in 2014/2015) onwards.

Statistical analysis

Descriptive statistics:

Unweighted descriptive statistics were reported as counts and percentages for categorical sociodemographic, pain status, and survival status variables at each wave. Note, some weighted pain prevalence statistics are available in the TILDA literature (Barrett et al., 2011).

Biases due to attrition and mortality:

An alluvial plot was used to visualise transitions in pain status (none, mild, moderate, severe), survival status, and lost to follow-up categories across the five waves.

Next, to investigate attrition bias, we fitted a logistic regression model of attrition (due to death or otherwise lost to follow-up) on pain severity. Attrition by Wave 5 was used as the outcome, with pain severity at Wave 1 as a predictor and controlling for Wave 1 sociodemographic factors. To investigate attrition bias due to mortality specifically, we fitted a logistic regression model of mortality by Wave 5 on pain severity, controlling for baseline sociodemographic factors. Those who were lost to follow-up by Wave 5 but not confirmed deceased were removed from this analysis. Adjusted odds ratios (AORs) with 95% confidence intervals (CIs) are reported for all logistic regression models. The likelihood ratio chi squared test statistic and McFadden's pseudo R squared (McFadden, 1979) are reported as measures of model fit.

Reporting heterogeneity:

Reporting heterogeneity is a form of measurement bias due to systematic differences in how distinct population subgroups self-report subjective conditions - in the case of our study, pain (Molina, 2016). Using logistic regression and Wave 1 data, we estimated the presence of reporting heterogeneity by examining differences across sociodemographic groups in the odds of reporting pain-related disability for a given severity of pain. Groups more likely to report pain-related disability at a given pain level may be experiencing pain that has higher impact, and thus may be "stoical" in their expression of pain severity. An AOR above 1 for a given sociodemographic group suggests such stoicism,

as members of this group are more likely to report pain-related disability for a given level of pain, compared to the reference group. This method of assessing differences in pain reporting styles is modelled on the approach in Grol-Prokopczyk (2017), but is exploratory rather than definitive; limitations of and alternatives to this approach are acknowledged in the Discussion. While not perfectly correlated, self-reported pain severity is strongly associated with pain-related disability in both clinical and population-based studies. A strong dose-response relationship has been observed whereby greater pain severity is predictive of greater disability (Covinsky et al., 2009), including for objective measures of function limitations such as gait speed (Simmonds et al., 2012).

Longitudinal sociodemographic disparities in pain:

To visually examine sociodemographic disparities in the reporting of pain, pain scores were calculated on the 0 to 3 scale, where 0 = no pain and 3 = severe pain. Differences in average pain scores across sociodemographic factors were then summarised at each wave using stratified line plots. Similarly, a line plot of average pain scores for survivors, decedents, those lost to follow-up, and the full sample across all waves illustrated if decedents had higher average pain scores than survivors. A multivariate latent growth curve model with pain severity status as an ordinal response (ordered categories: no pain, mild pain, moderate pain, severe pain) was fitted to estimate sociodemographic differences in pain trajectories. In brief, the model estimates the continuous latent response variable assumed to underlie the observed ordinal response variable. The latent pain trajectory is estimated in terms of an intercept and slope. Sociodemographic differences in pain trajectory are modelled as deviations from this intercept and slope. Thresholds are estimated to delineate the ordinal pain categories on the continuous latent variable scale. We assumed

longitudinal threshold invariance, meaning the thresholds for determining pain severity category from the latent variable were the same at each wave (Masyn et al., 2014). Root Mean Squared Error of Association (RMSEA) and Comparative Fit Index (CFI) are reported as measures of model fit. As a rule of thumb, RMSEA values below 0.08 and CFI values above 0.95 are considered indicative of “good fit” (Hooper et al., 2008).

All analyses were carried out in R (R Core Team, 2022). The `ggplot2` (Wickham, 2016) and `ggalluvial` (Brunson and Read, 2023) packages were used for data visualisation. The `lavaan` package was used for latent growth curve modelling (Rosseel, 2012). Base R and the `oddsratio` (Schratz, 2017) and `rcompanion` (Mangiafico, 2022) packages were used for the logistic regression models.

3.4 Results

Descriptive statistics

Table 3.1 details the number (%) of those remaining in the study at each follow-up wave across sociodemographic factors measured at Wave 1 and the recorded pain severity and pain-related disability across the waves. At Wave 1 60% of participants were aged 60 or over, with 7.7% of the sample aged 80 or over. Just over half (54.2%) the participants were women. Secondary was the most common highest level of education (39.9%), followed by none or primary (30.6%) and tertiary (29.4%). A majority (57.5%) reported having private health insurance.

By Wave 5, the proportion of the remaining sample who had reported private health insurance at Wave 1 was 64.2%. Of the Wave 5 sample 55.4% were aged 60 or over at Wave 1, while those aged 80 or over at Wave 1

Table 3.1: Descriptive statistics for Wave 1 sociodemographic factors and pain severity across waves in TILDA Waves 1-5.

Variables	Wave				
	1 (n = 8,171)	2 (n = 6,993)	3 (n = 6,246)	4 (n = 5,571)	5 (n = 4,872)
<i>Wave 1 age category</i>					
50-59	3,270 (40.0%)	2,862 (40.9%)	2,617 (41.9%)	2,402 (43.1%)	2,173 (44.6%)
60-69	2,593 (31.7%)	2,265 (32.4%)	2,077 (33.3%)	1,889 (33.9%)	1,694 (34.8%)
70-79	1,680 (20.6%)	1,394 (19.9%)	1,206 (19.3%)	1,034 (18.6%)	853 (17.5%)
80+	627 (7.7%)	472 (6.7%)	346 (5.5%)	246 (4.4%)	152 (3.1%)
<i>Sex</i>					
Male	3,743 (45.8%)	3,197 (45.7%)	2,833 (45.4%)	2,528 (45.4%)	2,191 (45.0%)
Female	4,428 (54.2%)	3,796 (54.3%)	3,413 (54.6%)	3,043 (54.6%)	2,681 (55.0%)
<i>Wave 1 education level</i>					
None or primary	2,503 (30.6%)	2,007 (28.7%)	1,685 (27.0%)	1,410 (25.3%)	1,135 (23.3%)
Secondary	3,262 (39.9%)	2,835 (40.5%)	2,563 (41.0%)	2,292 (41.1%)	2,026 (41.6%)
Tertiary	2,402 (29.4%)	2,150 (30.7%)	1,997 (32.0%)	1,868 (33.5%)	1,710 (35.1%)
<i>Wave 1 private health insurance status</i>					
Yes	4,702 (57.5%)	4,158 (59.5%)	3,838 (61.4%)	3,484 (62.5%)	3,129 (64.2%)
No	3,463 (42.4%)	2,832 (40.5%)	2,404 (38.5%)	2,084 (37.4%)	1,741 (35.7%)
<i>Pain severity at each wave</i>					
No pain	5,273 (64.5%)	4,553 (65.1%)	4,037 (64.6%)	3,634 (65.2%)	3,120 (64.0%)
Mild pain	829 (10.1%)	683 (9.8%)	551 (8.8%)	517 (9.3%)	459 (9.4%)
Moderate pain	1,349 (16.5%)	1,129 (16.1%)	1,042 (16.7%)	911 (16.4%)	833 (17.1%)
Severe pain	710 (8.7%)	545 (7.8%)	490 (7.8%)	404 (7.3%)	350 (7.2%)
<i>Pain-related disability at each wave¹</i>					
Yes	1,683 (58.2%)	1,373 (57.5%)	1,270 (59.4%)	1,165 (62.3%)	1,016 (60.2%)
No	1,210 (41.8%)	983 (41.1%)	816 (38.1%)	668 (35.7%)	630 (37.3%)

¹Percentages are for the number of participants who reported being often troubled by pain at each wave, not the entire sample.

represented just 3.1% of the Wave 5 sample. Secondary remained the most common highest Wave 1 education level by Wave 5 (41.6%), but tertiary became more common than none or primary education level (35.1% and 23.3% respectively). Just over half (55%) of the Wave 5 sample were women.

At Wave 1, 35.3% of participants reported they were often troubled by some degree of pain. The most commonly reported pain severity was moderate pain (16.5% of the full sample), followed by mild (10.1%) and then severe pain (8.7%). This distribution of pain severity remained very consistent across all five waves. This is evident in the latent growth curve model in Table 3.6 (discussed in more detail later), where the slope/rate of change in pain over time is not found to be significant. However, this finding may be affected

by attrition biases. These potential sources of bias are explored further in the following sections. The percentage of those with pain who reported pain-related disability was around 60% across waves. For those participating in any given wave, missing data for pain questions was low with item non-response at most 2.3%.

Table 3.2 summarises the percentage attrition at each wave, with a breakdown of what percentage of participants were confirmed deceased or otherwise lost to follow-up. 14.4% of the baseline sample did not return for Wave 2. This percentage increased to 23.6% at Wave 3 and 31.8% at Wave 4. By the beginning of Wave 5, a large proportion (3299, 40.4%) of baseline participants had left the study, either due to death or other loss to follow-up. 741 (9.1%) members of the baseline sample were confirmed deceased by this time, which accounted for 22.5% of the missing cases.

Table 3.2: Summary of attrition (deceased or otherwise lost to follow-up) at each wave.

Wave	Total attrition (out of n = 8,171)	Total confirmed deceased (out of n = 8,171)	Total otherwise lost to follow-up (out of n = 8,171)	% attrition who were confirmed deceased
2	1,178 (14.4%)	170 (2.1%)	1,008 (12.3%)	14.4%
3	1,925 (23.6%)	406 (5.0%)	1,519 (18.6%)	21.1%
4	2,600 (31.8%)	555 (6.8%)	2,045 (25.0%)	21.3%
5	3,299 (40.4%)	741 (9.1%)	2,558 (31.3%)	22.5%

Attrition and mortality bias:

Figure 3.1 is an alluvial plot showing the percentage in each pain category and transitions between pain categories across the five waves. Lost to follow-up (not confirmed deceased) and confirmed deceased categories are also included to capture sample attrition. The plot highlights the extent of mortality and lost to follow-up between waves. Of the remaining sample at each wave, the ratio of no pain to mild to moderate to severe pain appears consistent, as seen in Table 3.1.

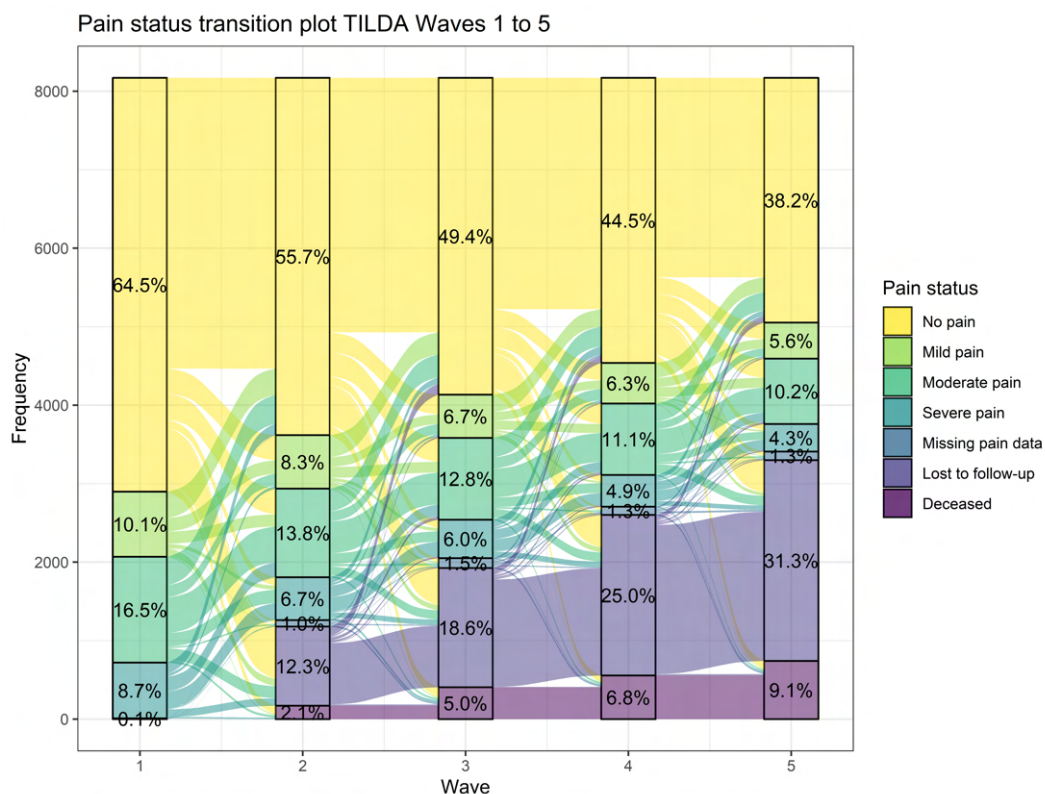


Figure 3.1: Transitions in pain status categories, including deceased and lost to follow-up, between waves. The thickness of the streams between pairs of categories at adjacent waves is proportional to the number of participants who transitioned between those two categories at those waves.

The alluvial plot also visualises pain category transitions over time. While there is considerable transition between categories, the majority of individuals at a particular wave tend to either stay in the same pain category or move to an adjacent pain category in the subsequent wave, excluding those who dropped out or died. For example, if we observe the severe pain category, the most common transitions between waves were to the same category or to the moderate pain category.

Results of the multivariable logistic regression of attrition (due to death or loss to follow-up) are presented in Table 3.3. Pain severity at Wave 1 was not a significant predictor of attrition by Wave 5, controlling for sociodemographic

factors. However, those with a tertiary level education at Wave 1 had considerably lower odds [AOR 0.48 (95% CI 0.42, 0.55)] of exiting the study by Wave 5 than those with none or primary education level, controlling for other factors including age. Those with secondary level education [AOR 0.67 (95% CI 0.59, 0.75)] also had lower odds of attrition. Not having private health insurance at Wave 1 was associated with increased odds of attrition by Wave 5 [AOR 1.45 (95% CI 1.32, 1.61)].

Table 3.3: Multivariable logistic regression of attrition (due to death or otherwise lost to follow-up) by Wave 5 on pain severity and sociodemographic factors at Wave 1 (n = 8,150)¹.

	AOR (95% CI)
Baseline sociodemographic characteristics	
<i>Age category (reference: 50-59)</i>	
60-69	0.96 (0.86, 1.07)
70-79	1.63 (1.44, 1.85)
80+	4.89 (4.01, 6.01)
<i>Sex (reference: Male)</i>	
Female	0.94 (0.85, 1.03)
<i>Education level (reference: None or primary)</i>	
Secondary	0.67 (0.59, 0.75)
Tertiary	0.48 (0.42, 0.55)
<i>Health insurance (reference: Yes)</i>	
No	1.45 (1.32, 1.61)
<i>Baseline pain severity (reference: No pain)</i>	
Mild pain	0.87 (0.74, 1.02)
Moderate pain	0.88 (0.78, 1.00)
Severe pain	1.10 (0.93, 1.30)
Measures of fit	
McFadden's R squared	0.07
Chi-squared test	795.06 (p<0.001)

¹21 observations were removed due to missingness on predictor variables.

Table 3.4 presents the results of the multivariable logistic regression of death by Wave 5, comparing those who were confirmed deceased with those

still participating at Wave 5, using pain severity at Wave 1 as a predictor while controlling for sociodemographic variables. Severe pain at Wave 1 was associated with a 63% increase in odds of death by Wave 5 [AOR 1.63 (95% CI 1.20, 2.19)] compared to those who were pain-free at Wave 1, controlling for age, sex, education level and health insurance status. Those with a secondary [AOR 0.70 (95% CI 0.56, 0.87)] or tertiary [AOR 0.63 (95% CI 0.48, 0.81)] level education at Wave 1 had lower odds of dying by Wave 5 than those with none or primary education level, while those without private health insurance had an estimated 75% higher odds of dying [AOR 1.75 (95% CI 1.43, 2.13)] compared to those with private health insurance.

Table 3.4: Multivariable logistic regression of death by Wave 5 on pain severity and sociodemographic factors at Wave 1 ($n = 5,600$)¹.

	AOR (95% CI)
Wave 1 sociodemographic characteristics	
<i>Age category (reference: 50-59)</i>	
60-69	2.55 (1.89, 3.47)
70-79	8.74 (6.60, 11.71)
80+	45.86 (33.31, 63.87)
<i>Sex (reference: Male)</i>	
Female	0.67 (0.56, 0.80)
<i>Education level (reference: None or primary)</i>	
Secondary	0.70 (0.56, 0.87)
Tertiary	0.63 (0.48, 0.81)
<i>Health insurance (reference: Yes)</i>	
No	1.75 (1.43, 2.13)
<i>Wave 1 pain severity (reference: No pain)</i>	
Mild pain	1.20 (0.89, 1.60)
Moderate pain	1.17 (0.91, 1.48)
Severe pain	1.63 (1.20, 2.19)
Measures of fit	
McFadden's R squared	0.26
Chi-squared test	1122.8 ($p < 0.001$)

¹13 observations were removed due to missingness on predictor variables. 2,558 Wave 1 participants who were lost to follow-up by Wave 5 but not confirmed deceased were removed.

Reporting heterogeneity measurement bias:

Differences in reporting pain-related disability between sociodemographic groups for given levels of pain severity at Wave 1 were examined as evidence of the presence of reporting heterogeneity using logistic regression, as given in Table 3.5. The analysis found the AOR for reporting pain-related disability of those without private health insurance cover was 1.37 (95% CI = [1.15, 1.64]), suggesting that these individuals were more impacted by experience of pain and may be more stoical in their expression of pain severity, compared to those with health insurance. Other sociodemographic characteristics were not significantly predictive of pain reporting style.

Longitudinal sociodemographic disparities in pain:

Figure 3.2 shows average pain scores in each wave stratified by Wave 1 sex, age category, highest education level, and health insurance status, and by survival status by Wave 5. In general, the trend lines are roughly flat, suggesting that average pain severity neither increases nor decreases over time within the different groups. However, average pain scores appear to vary considerably across the different sociodemographic groups.

Average pain scores were consistently higher for women than men at all waves of TILDA (Figure 3.2a). The average score for women was approximately 0.80 (on a scale from 0 to 3) in all waves, versus 0.55 for men. The differences in average pain scores between Wave 1 age categories across waves, Figure 3.2b, were small compared to the differences observed for sex. The youngest group (aged 50-59 at baseline) had the lowest average pain scores at approximately 0.65, followed by the 60-69 and then the 70-79 age groups. The average pain scores for the oldest group (aged 80+ at baseline) varied considerably, perhaps due to mortality between subsequent waves decreasing

Table 3.5: Multivariable logistic regression of pain-related disability on sociodemographic factors, controlling for pain severity, using Wave 1 data (n = 2,885)¹.

	AOR (95% CI)
Wave 1 sociodemographic characteristics	
<i>Age category (reference: 50-59)</i>	
60-69	1.06 (0.87, 1.28)
70-79	1.07 (0.86, 1.33)
80+	1.36 (0.99, 1.89)
<i>Sex (reference: Male)</i>	
Female	0.95 (0.80, 1.12)
<i>Education level (reference: None or primary)</i>	
Secondary	0.82 (0.67, 1.00)
Tertiary	0.86 (0.69, 1.09)
<i>Private health insurance (reference: Yes)</i>	
No	1.37 (1.15, 1.64)
<i>Wave 1 pain severity (reference: Mild pain)</i>	
Moderate pain	2.57 (2.15, 3.08)
Severe pain	8.80 (6.89, 11.32)
Measures of fit	
McFadden's R squared	0.11
Chi-squared test	430.44 (p<0.001)

¹8 observations were removed due to missingness on predictor variables.

the size of the group. Average pain scores for those aged 80+ at Wave 1 were the highest out of all the Wave 1 age categories, peaking at approximately 0.75.

There were clear and consistent disparities in pain scores across educational groups (Figure 3.2c). Those with no or primary levels of education at baseline consistently reported experiencing higher levels of pain, with average scores between approximately 0.75-0.85 across waves. Each subsequent increase in education level corresponded to lower average pain scores, approximately 0.65 for those with secondary level education and 0.55 for tertiary level education.

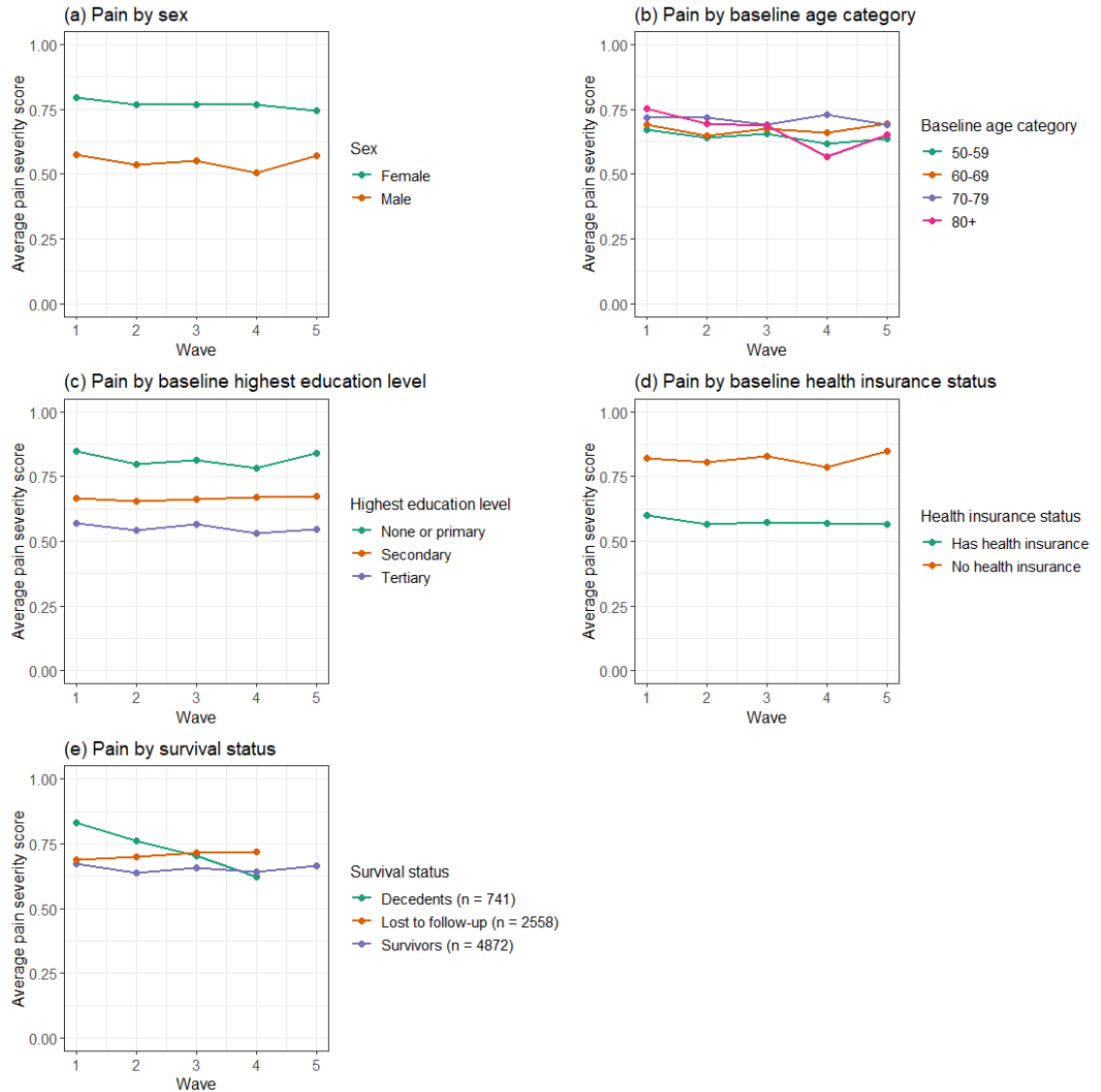


Figure 3.2: Average pain score over time by sociodemographic factors.

Figure 3.2d, which depicts average pain scores for those with and without private health insurance at baseline, reveals similar disparities in pain over time. The average pain score for those without private health insurance (approximately 0.85) is higher than for those with it (approximately 0.60) at all waves.

Results from the multivariate latent growth curve model of longitudinal sociodemographic disparities in pain are shown in Table 3.6. RMSEA = 0.015

and CFI = 0.998 indicate very good model fit. The estimated thresholds for the ordinal categories mild pain, moderate pain, and severe pain are 0.572, 0.870, and 1.649 respectively. The model parameters reflect the trends observed in the Figure 2 plots, including the relative flatness of the trend lines. All slope terms are not significantly different from zero, suggesting that propensity to experience pain is neither increasing nor decreasing overall or within the different sociodemographic groups over time.

Table 3.6: Multivariate ordinal latent growth curve model for pain severity status over 5 TILDA Waves.

	Intercept (95% CI)	Slope (95% CI)
<i>Wave 1 sociodemographic characteristics</i>		
Age category (reference: 50-59)		
60-69	-0.023 (-0.099, 0.053)	0.013 (-0.009, 0.035)
70-79	-0.069 (-0.165, 0.027)	0.026 (-0.001, 0.053)
80+	-0.160 (-0.356, 0.036)	0.025 (-0.034, 0.084)
Sex (reference: Male)		
Female	0.282 (0.213, 0.351)	-0.003 (-0.023, 0.017)
Education level (reference: None or primary)		
Secondary	-0.126 (-0.216, -0.036)	0.003 (-0.022, 0.028)
Tertiary	-0.228 (-0.326, -0.130)	-0.000 (-0.027, 0.027)
Private health insurance (reference: No)		
Yes	-0.239 (-0.313, -0.165)	-0.005 (-0.025, 0.015)
<i>Constant</i>	0.001 (-0.128, 0.130)	0.000 (-0.039, 0.039)
<i>Covariance of slope with intercept</i>	-0.045 (-0.055, -0.035)	
<i>Variances</i>	0.708 (0.679, 0.737)	0.024 (0.020, 0.028)
<i>Thresholds</i>		
No pain — Mild pain	0.572 (0.556, 0.588)	
Mild pain — Moderate pain	0.870 (0.856, 0.884)	
Moderate pain — Severe pain	1.649 (1.618, 1.662)	
<i>Measures of fit¹</i>		
RMSEA	0.015 (0.010, 0.020)	
CFI	0.998	

¹RMSEA = Root Mean Squared Error of Association, CFI = Comparative Fit Index.

However, intercepts did differ significantly by sex, education, and health insurance type. Women had a higher baseline propensity for pain [intercept 0.282 (95% CI 0.213, 0.351)] than men. Compared to those with no or primary level education, propensity for pain was successively lower for those with secondary [intercept -0.126 (95% CI -0.216, -0.036)] and tertiary [intercept

-0.228 (95% CI -0.326, -0.130)] level of education. Propensity for pain was also significantly lower for those with private health insurance at Wave 1 [intercept -0.239 (95% CI -0.313, -0.165)] than those without. Only the intercept terms for age categories were not significantly different to zero, suggesting no age disparities in pain. The variance of the latent variable intercept was significantly different to zero [0.708 (95% CI 0.679, 0.737)], indicating significant individual variation across participants in their propensity for pain. The variance in slope/rate of change in propensity for pain across participants was also significant [0.024 (95% CI 0.020, 0.028)]. The covariance between the latent variable intercept and slope suggested a weak inverse relationship between Wave 1 propensity for pain and rate of change over time [-0.045 (95% CI -0.055, -0.035)].

Figure 3.2e shows average pain scores for those who died during the 8-year Wave 1 to Wave 5 study period (decedents), as well as the average pain scores for those who left the study before Wave 5 but weren't confirmed deceased (lost to follow-up) and those who participated until Wave 5 (survivors). Average pain scores were highest for decedents at baseline then followed a negative linear trend, dropping below the average for those lost to follow-up or survivors by Wave 4. This suggests that those who had more severe pain at Wave 1 were more likely to die earlier in the study period and thus leave the study, inducing bias due to mortality. Survivors had the lowest average pain scores of the three groups up to Wave 4, and were consistently lower than the lost to follow-up group.

Figure 3.3 stratifies the decedent group by time-period of death. In general, the shorter the period of survival, the higher the average pain score at Wave 1 (with the exception of the "Died Wave 1 – Wave 2" group, whose baseline average pain score was very close to the "Died Wave 4 – Wave 5" group).

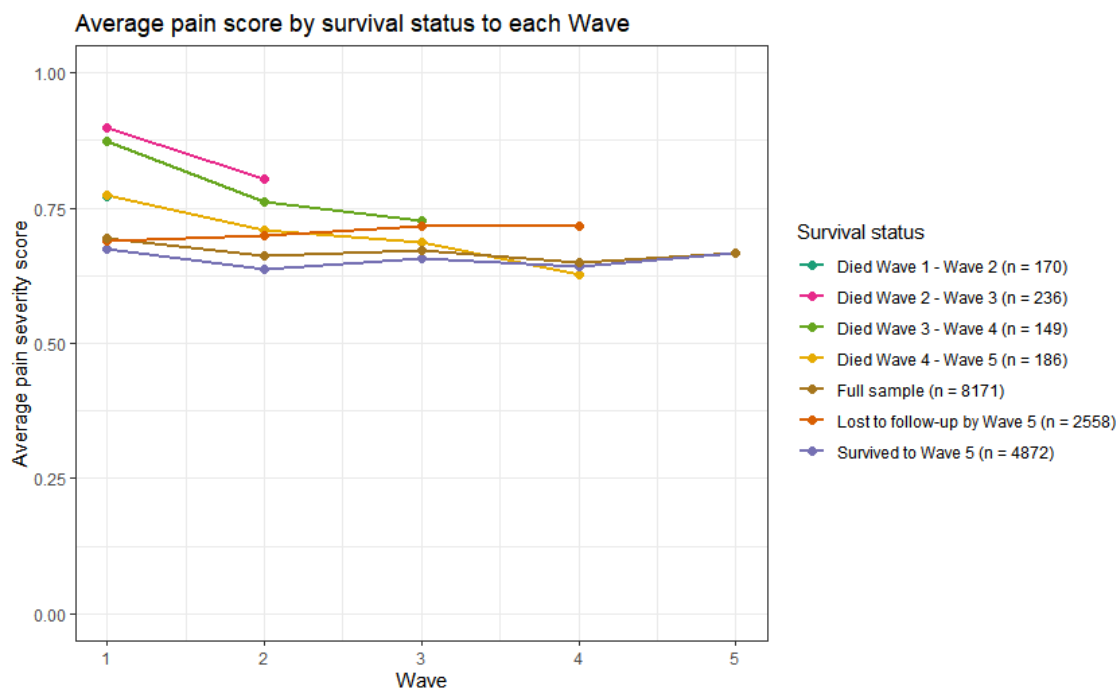


Figure 3.3: Average pain score over time by survival status.

We also note that average pain scores were consistently higher for those who were lost to follow-up than those who continued to participate to Wave 5. As participants who died or were otherwise lost to follow-up had higher mean pain scores than those who participated through Wave 5, those with higher pain levels become underrepresented in the sample over time. These findings suggest evidence of bias due to attrition.

3.5 Discussion and conclusions

Pain in older adults is a pervasive public health problem (Cohen et al., 2021) associated with negative outcomes including functional disability (Makino et al., 2019), frailty (Wade et al., 2017), reduced quality of life (Ludwig et al., 2018) and increased mortality risk (Torrance et al., 2010). Deriving accurate estimates of older adult pain prevalence and severity for different

sociodemographic groups is important to guide policy makers across Ireland and Europe in managing this problem. Our analysis suggests that the sex and socioeconomic disparities in pain found in other countries (Cimas et al., 2018; Grol-Prokopczyk, 2017; Ikeda et al., 2019; Lacey et al., 2013; Stewart Williams et al., 2015) are mirrored in Irish older adults. We also found evidence that Irish older adult pain estimates are subject to biases including reporting heterogeneity and mortality bias.

Attrition and mortality bias:

We found significant risk of pain-related mortality bias in TILDA, consistent with other longitudinal ageing studies. As those with more severe pain were more likely to die earlier in the study, there is a disproportionate loss of participants with severe pain. This mortality bias has serious implications. For example, studies of what factors predict pain will be biased if the data excludes people who already died and these people had higher-than-average pain. Additionally, this bias would lead to underestimation of the increase in pain with age in cross-sectional data, as people with more pain are more likely to die and so it will appear that pain risk does not rise with age. Pain severity was also a significant predictor of death in American (Grol-Prokopczyk, 2017) and Scottish older adult cohorts (Torrance et al., 2010). This combined evidence of pain-related mortality bias in older adult cohorts from different continents has important implications for pain research, requiring researchers to identify and mitigate such bias to avoid underestimating older adult pain experiences over time. Interestingly, while only severe pain was predictive of death in TILDA and the Scottish study, both moderate and severe pain were significant mortality predictors in the American analysis. This may suggest a stronger association between pain and death among American older adults.

We did not find statistically significant evidence of attrition bias in the TILDA cohort. However, while the effect estimates of pain severity on attrition were not large in either TILDA or the HRS (Grol-Prokopczyk, 2017), pain-related attrition bias should not necessarily be ruled out as a potential source of bias in older adult pain studies. A feature common to both TILDA and the HRS is that the samples consisted entirely of community-dwelling older adults at baseline, but participants who were later institutionalised were interviewed at subsequent waves when possible. As suggested in Grol-Prokopczyk (2017), studies that instead remove participants who move to care institutions from the follow-up sample may induce attrition bias, as participants with more severe pain may be lost. If present, attrition bias would result in the underestimation of pain prevalence.

Lower educational attainment and not having private health insurance were associated with increased odds of both mortality and attrition in TILDA, after controlling for age. These socioeconomic subgroups become underrepresented at later waves, an issue common in large voluntary studies of ageing (Brayne and Moffitt, 2022) which may induce bias. These findings highlight a need for initiatives to tackle socioeconomic barriers to sample recruitment and retention to reduce the risk of bias. Potential strategies include oversampling groups whose members are more likely to drop out, offering compensation or free travel to complete interviews and attend health assessments, and offering translated questionnaires for non-English speaking participants (Bonevski et al., 2014).

As some degree of attrition is typically unavoidable, we also highlight a need for awareness around the potential for such biases and methods to mitigate them, such as sample weighting. Multiple imputation (MI, Rubin, 1987) is one popular approach for handling missing data. However, standard MI methods assume the data is missing at random (MAR), meaning the missingness

depends only on the observed data (Van Buuren, 2018). As we found evidence that more severe pain is associated with mortality (missingness), it is plausible that the missingness is dependent on the missing (unobserved) pain values themselves. In this case, the data is missing not at random (MNAR), and results can be sensitive to violation of the MAR assumption (Carreras et al., 2021). Causal analysis designs such as instrumental variables (Tchetgen Tchetgen and Wirth, 2017) and inverse probability of censoring weighting (Rotnitzky et al., 1998) may be alternatives to handle this MNAR data. However, future research would be needed to examine the potential for these approaches to address the bias issues raised in this study.

Reporting heterogeneity:

We also found evidence of reporting heterogeneity, or measurement bias due to differences in self-reporting styles, at TILDA Wave 1. We interpret the results as those without private health insurance being more stoical in their reporting of pain than those with it. This result was similar to the multi-wave pooled analysis in Grol-Prokopczyk (2017), where participants with lower SES were found to be more stoical. While potential associations between lower SES and “good patient” behaviour have been posed previously (Pillay et al., 2014), there has been little work explicitly examining the extent and direction of reporting heterogeneity of persistent pain across sociodemographic groups. An Austrian study found a similar association between lower SES and disability while controlling for pain level, though the finding was labelled “unexplained” rather than attributed to possible reporting heterogeneity (Dorner et al., 2011). These findings may suggest a tendency in both Europe and America for socioeconomic pain disparities to be underestimated, due to those in less advantaged circumstances being stoical in their pain reporting. However,

unlike in the HRS study, women and those with no or only primary education were not found to be significantly more stoical in our TILDA analysis. This may suggest that the reporting habits of older adults and the propensity to be more stoical may vary by culture and geographical location. Determining the extent and direction of cross-group reporting heterogeneity should be a key consideration for any population study interested in group comparison, although strategies to overcome this type of bias, such as anchoring vignettes, are not straightforward (Grol-Prokopczyk et al., 2015).

Longitudinal sociodemographic disparities in pain:

Our results using Irish data confirm findings from other countries that the burden of pain in older adults is worse for women and those with lower SES (Jacobs et al., 2006; Milani et al., 2022; Palacios-Ceña et al., 2015; Wrangler et al., 2016). Average pain scores and propensity for pain were consistently higher for women compared to men across waves. Those without private health insurance and those with lower levels of education also had consistently higher average pain scores and propensity for pain than their higher SES counterparts across all waves. It is possible that age confounded some of the disadvantage for participants with a lower level of education in the trend plot, whereby those who were older appeared more likely to have a lower level of education and a higher risk of mortality. However, differences in propensity for pain across the education and health insurance categories remained significant after controlling for age in the multivariate latent growth curve model, while we found no significant evidence of age-related pain disparities in Irish older adults.

Some international trends were not reflected in the TILDA data. Research on other European countries has found that most (but not all) countries show a positive increase in pain prevalence over time, net of age (Zimmer et al.,

2020). Similar increasing pain trajectories over time have been reported in the US (Grol-Prokopczyk, 2017) and Canada (Shupler et al., 2019). In contrast, pain scores remained relatively flat across waves in TILDA and rates of change in propensity for pain were not significant. These differences may reflect genuinely different pain patterns among Irish older adults, perhaps due to differences in lifestyles and healthcare provision, for example. Alternatively, the flattened pain trend lines for TILDA may reflect the mortality bias suggested in our results.

Due to the relative homogeneity of the current older Irish population, we did not explore potential racial or ethnic pain disparities. However, we note that racial/ethnic disparities have been found in countries with more diverse older populations, such as the US (Morales and Yong, 2021). Such disparities may become relevant to Ireland’s policy makers as the diversity of the population increases.

A limitation of our study was the lack of an automated linkage system between survey data and death registration in Ireland. There is a time lag in the registration of deaths, so some participants who died may be categorised as “lost to follow-up” rather than “deceased” in our analyses. Without an objective measure of pain severity, evidence for the presence of reporting heterogeneity was sought as differences in pain-related disability experienced by groups with the same level of reported pain severity. This assumption that pain-related disability was reported without bias and only pain severity was subject to reporting heterogeneity is another limitation. The results of our reporting heterogeneity analysis are therefore exploratory. Future work is required to validate these findings and further examine differences in pain reporting styles. This could be done by comparing pain self-reports to objectively timed walking speeds, which are highly correlated with pain

severity (Hicks et al., 2017; Simmonds et al., 2012), or to measures such as frailty indices, which have been used to assess discrepancies in self-rated health (Calvey et al., 2022). The measures of pain prevalence (“are you often troubled by pain?”) and pain severity (“How bad is the pain most of the time?”) are also somewhat limited, as is the dichotomous indicator of pain-related disability, which does not convey degree of disability. However, these pain questions are used across multiple global ageing cohort studies, which has the benefit of allowing direct replication and comparison of our results across ageing populations in different countries (Gateway to Global Aging, 2023). Finally, reporting heterogeneity was explored using Wave 1 data only. Future research could investigate changes in reporting heterogeneity over time.

A key message from this work is that the potential for bias in population studies cannot be ignored. Failure to investigate and account for such biases may result in inaccurate estimates of pain prevalence and pain disparities in older adult populations, weakening the evidence base which guides policy makers’ decisions. Identifying bias is also an important step for potential future work looking at causal relationships between pain and attrition. Additionally, we highlight a need to address sociodemographic disparities in pain among Irish older adults. Targeted interventions are required to tackle the disproportionate pain burden of women, those with lower levels of education and without private health insurance.

Many countries across Europe and the world have ageing populations, which present economic and healthcare challenges (Christensen et al., 2009). Accurate estimates of pain prevalence, along with an understanding of pain trajectories and sociodemographic disparities in pain will be required for policy makers and health services to plan appropriately. This work is a first step towards providing such estimates for the older Irish population, while

highlighting biases that may impact pain research using observational studies in Ireland and internationally.

3.6 Acknowledgements

The authors gratefully acknowledge the participants in the TILDA study, the study nurses, administrators, and members of the TILDA research team. TILDA data was accessed via the TILDA hot desk facility on the TCD campus. Researchers seeking access to the full TILDA dataset may apply to access the data on the TCD campus (tilda.tcd.ie). Applications are considered on a case-by-case basis. All R code employed in this paper will be made available to applicants on request.

3.7 Ethics approval

The TILDA study received ethical approval from the Faculty of Health Sciences Research Ethics Committee at Trinity College Dublin.

3.8 Author contributions

ER, AH and HP conceptualised the design of the study. ER conducted the analyses, interpreted the results and wrote the first draft of the manuscript. PM supported access to the data. All authors interpreted the results, critically revised the manuscript, approved the final version, and agreed with its submission to the European Journal of Pain.

3.9 Funding information

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049. Research reported in this publication was also supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG065351 (PI: Grol-Prokopczyk). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4 Is the Relationship Between Chronic Pain and Mortality Causal? A Propensity Score Analysis

The work presented in this chapter has been published in PAIN[®]:

Ryan, E., Grol-Prokopczyk, H., Dennison, C., Zajacova, A., & Zimmer, Z. (2024). Is the relationship between chronic pain and mortality causal? A propensity score analysis. *PAIN*[®] (Accepted and published online: <https://doi.org/10.1097/j.pain.0000000000003336>).

The content of the article is reprinted verbatim with minor formatting differences. Supplementary materials provided in an online appendix have been included in the final section of the chapter for ease of reference. The author of this thesis (ER) conceptualised the design of this study along with HG-P, AZ, and ZZ. ER was also responsible for designing and carrying out the analysis, writing the first draft of the manuscript, and editing the manuscript. All authors interpreted the results, critically revised the manuscript, and approved the final version of the manuscript. This work was initiated during an international research visit as part of ER's Ph.D. programme.

4.1 Abstract

Chronic pain is a serious and prevalent condition that can affect many facets of life. However, uncertainty remains regarding the strength of the association between chronic pain and death and whether the association is causal. We investigate the pain–mortality relationship using data from 19,971 participants aged 51+ years in the 1998 wave of the U.S. Health and Retirement Study. Propensity score matching and inverse probability weighting are combined with Cox proportional hazards models to investigate whether exposure to chronic pain (moderate or severe) has a causal effect on mortality over a 20-year follow-up period. Hazard ratios (HRs) with 95% confidence intervals (CIs) are reported. Before adjusting for confounding, we find a strong association between chronic pain and mortality (HR: 1.32, 95% CI: 1.26-1.38). After adjusting for confounding by sociodemographic and health variables using a range of propensity score methods, the estimated increase in mortality hazard caused by pain is more modest (5%-9%) and the results are often also compatible with no causal effect (95% CIs for HRs narrowly contain 1.0). This attenuation highlights the role of confounders of the pain–mortality relationship as potentially modifiable upstream risk factors for mortality. Posing the depressive symptoms variable as a mediator rather than a confounder of the pain–mortality relationship resulted in stronger evidence of a modest causal effect of pain on mortality (e.g., HR: 1.08, 95% CI: 1.01-1.15). Future work is required to model exposure–confounder feedback loops and investigate the potentially cumulative causal effect of chronic pain at multiple time points on mortality.

4.2 Introduction

Chronic pain (sometimes abbreviated as “pain” herein) is a potentially debilitating condition affecting an estimated 20% of American adults (Dahlhamer, 2018). Pain is known to impact many facets of life, including quality of life (Hadi et al., 2019; Leadley et al., 2014), psychological well-being (Atkinson et al., 1991), work productivity (Blyth et al., 2003; Kawai et al., 2017), and social functioning (Dueñas et al., 2016). Whether chronic pain affects not only the quality of life but also the quantity of life, however, is less clear.

Net of spurious associations due to confounding, one might expect pain to influence mortality for a number of reasons. Pain may reduce mobility (Makris et al., 2014) and functional ability (Makris et al., 2017), which in turn may have detrimental effects on both physical and mental health (Froehlich-Grobe et al., 2016; Musich et al., 2018), leading to greater mortality risk. Another plausible causal pathway is through the use of opioid analgesics for pain management. Opioids have been shown to increase mortality through multiple mechanisms, especially cardiovascular deaths (Ray et al., 2016; Tölle et al., 2021).

Empirical research on chronic pain and mortality is characterized by a number of unresolved issues. One is the degree to which pain and mortality are associated net of confounding variables. There is a high degree of heterogeneity in observational studies of this association (Smith et al., 2014), with some finding a significant positive association after adjusting for potential confounders (Macfarlane et al., 2001; McBeth et al., 2009; Nitter and Forseth, 2013; Sjøgren et al., 2009) and others finding no association (Andersson, 2009; Dreyer et al., 2010; Macfarlane et al., 2007; Smith et al., 2003; Torrance et al., 2010; Wolfe et al., 2011). This discordance may be attributable to methodological differences, such as different follow-up lengths, population

characteristics, methods of analysis, confounder adjustments (Smith et al., 2014), or pain phenotypes (Smith et al., 2018). Other possible explanations may be contextual, such as different healthcare systems across countries (Melchiorre et al., 2013; Van Eeⁿoo et al., 2016).

A second unsettled issue—and a more contentious one—is the degree to which any association between pain and mortality is causal. Randomised control trials remain the “gold standard” of causal research, but conducting trials of long-term pain exposure is infeasible and unethical. This necessitates the analysis of observational data. However, robust methods specifically designed for causal analysis of observational data (Hernán and Robins, 2020) have rarely been applied in pain/mortality research. Many existing studies do not apply an explicitly causal approach but, as is often the case in observational research, use language that could be interpreted as agnostic to causality or as implying causality (Haber et al., 2022).

A handful of studies explicitly address the causal relationship between pain and mortality. An analysis of U.S. National Health and Nutrition Examination Survey data found that pain had a causal effect on 3- and 5-year mortality mediated by opioid prescriptions (Inoue et al., 2022). Pain was also found to affect mortality in an analysis of the English Longitudinal Study of Ageing (ELSA; Smith et al., 2018). This prior study adjusted for age, sex, education, and wealth, but treated all other variables (multiple health, lifestyle, social and psychological factors) as mediators.

Whether pain has a causal effect on death or is only associated with death due to confounding factors has important implications for the design of interventions to reduce mortality. The aim of this study is to expand our understanding of the pain–mortality relationship by using propensity score methods to explicitly address the question: “Does pain have a causal effect on

mortality among older American adults?”

4.3 Methods

Data

This study uses data from the Health and Retirement Study (HRS; Health and Retirement Study, 2023). The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. Detailed documentation of the study design and data collection is available on the HRS website (Institute for Social Research University of Michigan, 2023). In brief, the HRS uses a multistage area probability sampling design with geographical stratification and clustering. African-American and Hispanic adults are oversampled. Sampling weights are provided in the HRS datasets to account for differential selection probabilities. The HRS began in 1992 with a cohort of 50- to 60-year-old respondents. The Asset and Health Dynamics Among the Oldest Old study was established 1 year later to study a cohort aged 70 years and older. By 1998, the HRS and Asset and Health Dynamics Among the Oldest Old cohorts were combined along with 2 new birth cohorts to fill the age gaps, making the HRS a nationally representative sample of non-institutionalised adults aged over 50 years in the United States (Sonnegga et al., 2014). Study participants are followed up every 2 years by telephone or in-person interviews. This study analyses community-dwelling adults aged 51 years and above from the 1998 HRS sample. It follows this sample over a 20-year period from 1998 to the end of 2018, monitoring their mortality over this period. Of the 20,003 participants aged 51+ years at baseline in 1998, 32 were excluded due to missing pain data, yielding an analytic sample of 19,971 participants.

Variables*Exposure:*

The 1998 wave of the HRS included multiple questions on pain experience. Participants were first asked, “Are you often troubled by pain?” This question is believed to capture persistent rather than transient or trivial pain (Grol-Prokopczyk, 2017), with previous research finding that study participants reported being “often troubled by pain” less than half as often as they reported “any pain in the last 30 days” (Banks et al., 2009). Those who responded yes to the first pain question were then asked, “How bad is the pain most of the time: mild, moderate, or severe?” We defined pain “exposure” as reporting moderate or severe pain at the 1998 wave, with no pain or mild pain as a control. We chose these groupings because prior research has found that moderate and severe pain, but not mild pain, are significant predictors of mortality in older adults (Grol-Prokopczyk, 2017; Smith et al., 2018). Sensitivity analyses comparing severe pain exposure to none/mild/moderate pain, comparing any pain exposure to no pain, and comparing just severe pain to no pain were conducted and gave similar results to the severe/moderate vs none/mild pain groupings (see Appendix 2 for supplemental results, <http://links.lww.com/PAIN/C93>).

Outcome:

The outcome of interest is survival over 252 months from January 1998 (beginning of the 1998 HRS survey year) to December 2018 (end of the 2018 survey year). HRS mortality data are essentially complete. Mortality information is collected in 2 ways: either from family members when attempting to contact a participant for the next survey wave, or through linkage to the National Death Index (Weir, 2016). The month and year of

death are reported for deceased participants.

Confounders:

Careful consideration of both potential confounders and appropriate methods to adjust for them is key to drawing causal inferences from observational data (Hernán and Robins, 2020). This decision cannot be entirely data driven and must use expert knowledge, as statistical associations alone generally cannot distinguish between mediators (which should not be adjusted for) and confounders (which should be) (VanderWeele, 2019). The biopsychosocial model of pain adapted for older adults suggests that chronic pain experience is influenced by a range of factors, including biological (e.g., sex, age, body mass index [BMI], smoking status, health conditions), psychological (e.g., depression), and social (e.g., race, socioeconomic status [SES], social isolation) ones (Miaskowski et al., 2020). Similar biological, psychological, and social factors have also been associated with differential mortality (Elo, 2009; Holt-Lunstad et al., 2015; Olshansky et al., 2012; Wang et al., 2012; White et al., 2016). Thus, the relationship between pain and mortality may be conceptualised as a complex causal framework encompassing demographic characteristics, socioeconomic factors, health behaviours, psychological factors, and medical conditions (Zajacova et al., 2021). A comprehensive confounder adjustment set spanning each of these categories was chosen a priori for this analysis based on the existing pain and mortality literature, as detailed below.

There is evidence in the literature of demographic disparities in chronic pain (Grol-Prokopczyk, 2017; Mullins et al., 2022), with characteristics including increased age, female sex/gender, being divorced or separated, and geographic location identified as risk factors (Mills et al., 2019; Sjøgren et al., 2009; Van Hecke et al., 2013; Zajacova et al., 2022). Similar demographic disparities

in life expectancies and mortality have also been observed (Holt-Lunstad et al., 2015; Olshansky et al., 2012; Spencer et al., 2018; Wang et al., 2012). Demographic variables included as potential confounders in this analysis were age in years in 1998; sex (male, female); race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, non-Hispanic Other); marital status (married, separated/divorced, widowed, never married, other); household size; number of children; region of residence (Northeast, Midwest, South, West); and urbanicity (urban, suburban, exurban/rural).

In addition, it is well established that chronic pain is inversely associated with SES (Mills et al., 2019). Indicators of lower SES such as less wealth and lower educational attainment have also been associated with increased mortality risk (Davies et al., 2018; Elo, 2009). SES variables included as potential confounders in this study were the highest level of education (no degree, high school diploma, 4-year college degree, graduate degree); household wealth quartile; employment status (employed, unemployed, retired, not in labour force); food security (“In the last 2 years, have you always had enough money to buy the food you need?” yes, no); veteran status (yes, no); and health insurance type (uninsured, any private insurance, public insurance only). We also included a variable about the importance of religion (very important, somewhat important, not too important), as the centrality of religion in individuals’ lives has been associated with both differential pain experience (Dezutter et al., 2010) and mortality (Idler et al., 2017).

A range of chronic conditions including diabetes, cancer, cardiovascular conditions, and depression have been associated with greater pain burden in older adults (Kroenke et al., 2011; Parsons et al., 2015; Sjøgren et al., 2009; Zajacova et al., 2021). These conditions are also well recognized as causes of death (Crimmins and Beltrán-Sánchez, 2011; Kvale et al., 2011; Ma et al.,

2015). Other key health factors known to be associated with both pain and mortality include BMI (Flegal et al., 2013; Qian et al., 2021; Ray et al., 2011; Tobias and Hu, 2018) and smoking (Gellert et al., 2012; Mills et al., 2019). Health status variables included as potential confounders in this study were active cancer (yes, no); diabetes (yes, no); chronic lung disease (yes, no); angina (yes, no); stroke (yes, no); heart condition (yes, no); arthritis (yes, no); BMI category (underweight BMI < 18.5, normal weight $18.5 \leq \text{BMI} < 25$, overweight $25 \leq \text{BMI} < 30$, obese 1 $30 \leq \text{BMI} < 35$, obese 2 $35 \leq \text{BMI} < 40$, obese 3 BMI 40+); smoking status (never smoker, former smoker, current smoker); and mental health status as measured by the 8-item version of the Center for Epidemiological Studies-Depression Scale (0-8, with higher scores indicating more depressive symptoms) (Karim et al., 2015).

We note that the association between chronic pain and depression is well established (Bair et al., 2003). However, the direction of the relationship between these conditions is complex and likely reciprocal (Bondesson et al., 2018; Cohen et al., 2021), partly due to shared neural mechanisms (Hooten, 2016). Longitudinal studies have found that depression is an upstream risk factor for pain onset (Carroll et al., 2004; Currie and Wang, 2005), and we assume this is the case in our main analyses. On the other hand, pain may also increase depression risk (Sheng et al., 2017), for example, as a maladaptive response to pain-related disability (Surah et al., 2014). The potentially bidirectional nature of the pain–depression relationship complicates our investigation of the effect of pain on mortality, as it is unclear if depressive symptoms should be included as a cause or as a result of pain in the causal model (i.e., as confounder or mediator in our analyses). As detailed later in this section, we contend with this ambiguity by repeating our analyses with depressive symptoms treated as a mediator rather than a confounder.

Analysis

As detailed in the following subsections, we use propensity score approaches and Cox proportional hazard models to analyze the data. Missing data are handled using multiple imputations. An overview of the analytical process is provided in Figure 4.1.

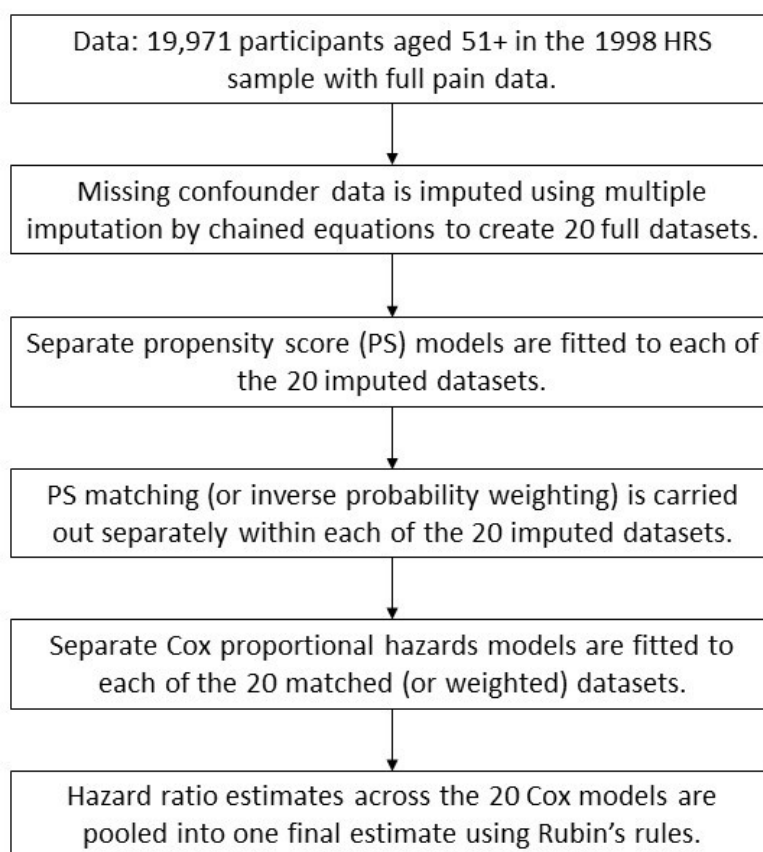


Figure 4.1: Flowchart summarizing the analytic procedure.

Missingness and imputation:

There was a small amount of missingness on the 1998 variables chosen for inclusion as covariates in the propensity score model. As adult height is quite consistent over time, missing height data needed to calculate the

BMI categories were first propagated forward using a last observation carried forward approach from the first HRS wave in 1992. Height values starting from the 2018 wave were then propagated backwards to fill some of the remaining missing height values. After this step, the remaining missing values across all 1998 covariates were imputed using multiple imputations by chained equations (Van Buuren, 2018). Variables included in the analyses were used in the imputation models. Propensity score methods and Cox proportional hazard models were applied separately to each of the 20 imputed datasets. The results were then pooled across imputations using Rubin’s rules, an approach designed to account for both within-imputation and between-imputation variance (Rubin, 1987).

Propensity score approach:

A key assumption for causal interpretation is exchangeability, meaning that the pain-exposed and -unexposed groups have similar distributions of background characteristics and differ only by exposure. This ensures that differences in outcomes between groups can be solely attributed to pain exposure. Exchangeability is achieved in controlled trials by randomization but can also be achieved in observational studies if the conditional probability of exposure depends only on measured covariates (confounders), meaning there is no unmeasured confounding (Hernán and Robins, 2020).

Propensity score methods are used to achieve balance on measured confounders across exposed and unexposed groups. The propensity score is the probability of exposure conditional on observed background characteristics. Conditional on the propensity score, the distribution of these background characteristics will be similar in the exposed and unexposed groups (Austin, 2011). For our propensity score model, we fit a multiple logistic regression

model with pain exposure as the dependent variable and demographic, socioeconomic, and health variables as predictor variables. This model was then used to predict the probability of pain exposure for each participant, and these probabilities were used as propensity scores.

After calculating the propensity scores, we used 2 different methods to achieve balance across the exposed and unexposed groups: propensity score matching and inverse probability weighting (IPW). Propensity score matching involves creating a matched sample by pairing pain-exposed participants with unexposed participants based on the similarity of their scores. Our primary matching approach, often found to outperform other approaches (see Appendix 1 in Section 4.7), was one-to-one nearest neighbour caliper matching without replacement. Inverse probability weighting involves weighting the sample such that the pain-exposed and -unexposed groups have similar distributions of background characteristics. We also used doubly robust models and regression-adjusted matching, which account for confounding both in the propensity score model and in the survival model for the outcome. Finally, as there are many different approaches to propensity score matching, we conducted additional analyses using alternative matching approaches. These were 1-to-1 matching with replacement, 2-to-1 matching with replacement and without replacement, and optimal full matching. Details of all propensity score approaches are included in Appendix 1 (Section 4.7).

After applying propensity score methods, it is important to perform balance diagnostics to assess whether sufficient balance on background characteristics between the exposed and unexposed groups has been achieved (Austin, 2009a). Balance is typically assessed by comparing the difference in means and proportions for the various background variables between the exposed and unexposed groups. It is advisable to calculate standardized differences, as they

allow variables of different scales to be compared and they are not influenced by sample size (Ali et al., 2014; Austin and Stuart, 2015). As a general rule, absolute standardized differences below 0.10 are considered indicative of good balance (Austin, 2009b, 2011). For this study, the mean and range of standardized absolute mean differences across imputations are summarized in a covariate balance plot, with a dashed vertical line marking the 0.10 absolute standardized mean difference (ASMD) cut-off.

The goal of propensity score methods is to adjust for all possible confounders such that the causal effect of the exposure on the outcome can be identified. When selecting the confounder adjustment set to include in the propensity score model, it is important to avoid adjusting for intermediate variables on the causal pathways between exposure and outcome (mediators of the causal effect). Such “over-adjustment” can introduce serious bias (Van Zwieten et al., 2022; VanderWeele, 2009b). So-called “collider” selection bias could also be introduced by adjusting for variables that are not confounders, depending on their position in the underlying causal structure (Mansournia et al., 2013). Causal directed acyclic graphs (DAGs) are useful tools to visualize the exposure and outcome of interest along with all related variables and the proposed causal pathways between them, aiding the appropriate identification of confounders (Tennant et al., 2021). We include a DAG to provide such a visualization.

For some variables, determining their likely position in the causal framework was difficult. Specifically, as discussed in more detail in the Confounders section, pain and depression are believed to have a reciprocal relationship (Bondesson et al., 2018) wherein pain may also cause depression (Sheng et al., 2017). Thus, depressive symptoms could be a confounder or mediator of the pain–mortality relationship depending on temporal order. To assess whether treating depressive symptoms as a mediator rather than a confounder

would affect our results, all analyses were repeated without adjusting for depressive symptoms in the propensity score models, regression-adjusted matching models, or doubly robust IPW models.

Cox proportional hazard models:

Cox proportional hazard models with pain exposure as a covariate were fitted to the propensity score matched samples and to the IPW samples to estimate the hazard ratio (HR) of pain exposure vs nonexposure. Cox models are a widely used approach for survival analysis (Kleinbaum and Klein, 1996) that can be combined with propensity score methods in causal analyses (Austin, 2014b). Pain exposure HRs and 95% confidence intervals (CIs) for the HRs are reported for all models. The estimated HRs are interpreted as the instantaneous mortality rate at any time during the follow-up for those who were exposed to pain compared with those who were not (Sutradhar and Austin, 2018). Participants who were lost to follow-up during the study period were right censored in January of the year of the first wave in which they did not participate, e.g., someone who participated in 1998 but did not return for the next wave in 2000 was censored in January 2000. We also fitted an unadjusted Cox proportional hazards model by pain in 1998 to the original imputations before propensity score methods were applied. We then repeated this analysis while adjusting for age, age squared, and sex as covariates. The results of these models were then compared with the analyses using propensity score methods.

Kaplan–Meier curves:

Kaplan–Meier curve plots are used to visualize survival over time stratified by pain exposure in the unadjusted data ($n = 19,971$) and the 20 imputed and

one-to-one matched without replacement datasets. Mean survival probabilities were pooled across the 20 datasets for the matched data Kaplan–Meier curves. To our knowledge, there is no accepted convention for pooling variances across imputed and matched datasets for Kaplan–Meier plots, so CIs were not calculated for the matched data plot. All analyses were carried out in R (R Core Team, 2022) using the following packages: tidyverse (data preprocessing; Wickham et al., 2019); mice (missing data imputation; Van Buuren and Groothuis-Oudshoorn, 2011); MatchThem (PS matching and IPW of the imputed datasets; Pishgar et al., 2021); survival (survival models; Therneau, 2022); survminer (Kaplan–Meier curve plots; Kassambara et al., 2021); and cobalt (balance plots; Greifer, 2023). R code is available in the following GitHub repository: <https://github.com/Eva-Ryan/hrs-pain-mortality>.

Sensitivity analyses:

We conducted multiple sensitivity analyses to test the reliability of our findings.

1. All models in our primary analysis were fitted without using HRS sample weights, similar to the approaches taken in previous studies of the causal effect of pain on mortality (Inoue et al., 2022; Smith et al., 2018). As a sensitivity analysis, we re-ran all models with sample weighting (results summarized in Section 4.8 Appendix 2, Supplementary Table A4.1). Some research has found that incorporating sample weights in both the propensity score model and the outcome model is the most robust approach (Ridgeway et al., 2015). Thus, we first applied the sample weights when fitting the logistic regression models used to calculate the propensity scores. Then, after propensity score matching, the sample weights were applied when fitting the Cox proportional hazards outcome

models to the matched samples. For the IPW analyses, the propensity score weights were first multiplied by the sample weights to create new weights, and these composite weights were then applied when fitting the outcome models (DuGoff et al., 2014).

2. We explored the sensitivity of our confounder selection, specifically the inclusion of arthritis as a confounder, since arthritis may not confound the pain–mortality relationship. Similar to the alternative analysis in the main text where depressive symptoms were removed as a confounder, we repeated the analyses without adjusting for arthritis as a confounder (results summarized in Section 4.8 Appendix 2, Supplementary Table A4.2).
3. As noted earlier, we explored different pain specifications to test the sensitivity of our pain exposure groupings (pain exposed = moderate or severe pain, not pain exposed = no or mild pain). The analyses were repeated using the alternative groupings severe pain exposure (severe pain) vs no severe pain exposure (no, mild, or moderate pain), any pain exposure (mild, moderate, or severe pain) vs no-pain exposure (no pain), and severe pain exposure (severe pain) vs no-pain exposure (no pain). The results are summarized in Section 4.8 Appendix 2, Supplementary Tables A4.3, A4.4, and A4.5 and respectively).
4. We explored shorter follow-up lengths to test if the effect of pain exposure in 1998 on subsequent mortality is weakened over time. The analyses were repeated for 1-, 5-, and 10-year follow-up periods. The results are summarized in Section 4.8 Appendix 2, Supplementary Table A4.6.
5. To provide reassurance that our defined pain exposure is likely to be capturing persistent rather than transient pain, we re-ran our analyses with our exposure group restricted to “moderate/severe pain AND

arthritis” with “none/mild pain” as the comparison group as in the main analysis. We reason that if a person reports both pain and arthritis, that person is likely to be referring to chronic pain (resulting from their arthritis). We suggest that if the results are similar to our original analysis, then our original exposure definition likely also refers primarily to persistent pain. The results are summarized in Section 4.8 Appendix 2, Supplementary Table A4.7.

4.4 Results

The assumed causal structure underlying our analysis is shown in the DAG in Figure 4.2. The exposure of interest (pain) is shown in green, the outcome of interest (mortality) is in blue, and confounders of the pain–mortality relationship are shown in white boxes. Directed arrows depict the assumed direction of causality between variables. The absence of an arrow between any 2 boxes indicates no assumed causal relationship.

Table 4.1 describes potential confounding variables for both the pain ($n = 4,073$) and no-pain ($n = 15,898$) groups in the imputed datasets before propensity score methods are applied. Means/proportions are averaged across the 20 imputations. Covariate balance before applying propensity score methods is also summarized as averaged standardized mean differences across the imputations. There is an imbalance between the pain and no-pain groups on many background characteristics, as indicated by $ASMD > 0.10$ (in bold in Table 4.1). The largest imbalances are for arthritis ($ASMD = 0.874$), depressive symptoms ($ASMD = 0.608$), angina ($ASMD = 0.287$), heart condition ($ASMD = 0.261$), being female ($ASMD = 0.235$), and not being in the labour force ($ASMD = 0.400$), which are all more common in the pain

group, and being employed (ASMD = 0.373), being in the top wealth quartile (ASMD = 0.272), and having a graduate degree (ASMD = 0.256), which are more common in the no-pain group.

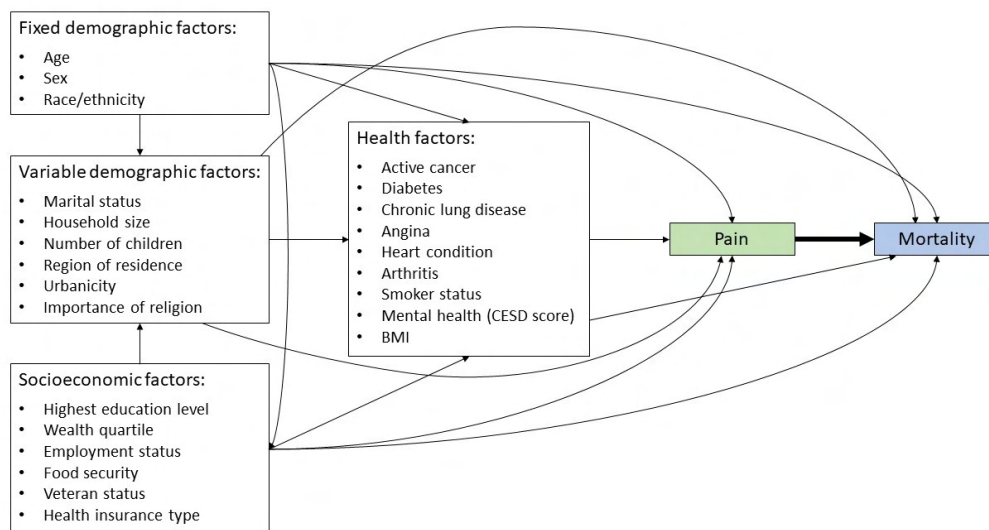


Figure 4.2: Directed acyclic graph showing assumed causal structure underlying the effect of pain (exposure) on mortality (outcome). Individual variables are grouped into “super nodes” in the DAG to improve readability. Some relationships between confounder variables have been simplified as a result (e.g., while a causal arrow goes from “Socioeconomic factors” to “Variable demographic factors,” the individual demographic factor “Number of children” could causally affect the individual socioeconomic factor “Wealth quartile”). However, this simplification should not affect how well the selected confounder adjustment set accounts for confounding of the effect of pain on mortality. It is likely that a number of factors, such as functional ability or opioid analgesic use, mediate the causal effect of chronic pain on mortality. However, as we do not conduct any mediation analyses in this study, we have not included potential mediators of the pain–mortality relationship in our DAG.

Table 4.1: Baseline characteristics for individuals with and without pain averaged across the 20 imputed data sets.

Variable	Averages for imputed datasets (before propensity score methods applied) (n = 19,971)		
	Pain (Exposed) (n = 4,073)	No Pain (Unexposed) (n = 15,898)	Standardized Mean Differences
<i>Age in years</i>	67.40	67.00	0.05
<i>Sex: Female</i>	0.66	0.54	0.24
<i>Race/ethnicity</i>			
White (non-Hispanic)	0.75	0.77	-0.05
Black (non-Hispanic)	0.14	0.14	0.02
Hispanic	0.09	0.08	0.06
Other (non-Hispanic)	0.02	0.02	-0.01
<i>Marital status</i>			
Married	0.62	0.66	-0.10
Separated/divorced	0.13	0.11	0.08
Widowed	0.22	0.20	0.05
Never married	0.03	0.03	0.00
Other	0.01	0.00	0.02
<i>Household size</i>	2.28	2.25	0.02
<i>Number of children</i>	3.27	3.17	0.05
<i>Region</i>			
Northeast	0.16	0.17	-0.02

Variable	Averages for imputed datasets (before propensity score methods applied) (n = 19,971)		
	Pain (Exposed) (n = 4,073)	No Pain (Unexposed) (n = 15,898)	Standardized Mean Differences
Mid-west	0.24	0.25	-0.03
South	0.42	0.41	0.02
West	0.18	0.17	0.03
<i>Urbanicity</i>			
Urban	0.46	0.49	-0.06
Suburban	0.22	0.22	0.00
Ex-urban/rural	0.32	0.30	0.06
<i>Religion</i>			
Very important	0.68	0.64	0.09
Somewhat important	0.23	0.25	-0.06
Not too important	0.10	0.11	-0.06
<i>Education level</i>			
No degree	0.36	0.27	0.20
High school degree	0.54	0.55	-0.02
4-year college degree	0.07	0.11	-0.16
Graduate degree	0.03	0.08	-0.26
<i>Wealth quartile</i>			
Q1	0.35	0.22	0.27
Q2	0.25	0.25	0.01
Q3	0.21	0.26	-0.12
Q4 (wealthiest)	0.18	0.27	-0.22

Variable	Averages for imputed datasets (before propensity score methods applied) (n = 19,971)		
	Pain (Exposed) (n = 4,073)	No Pain (Unexposed) (n = 15,898)	Standardized Mean Differences
<i>Employment status</i>			
Employed	0.22	0.37	-0.37
Unemployed	0.02	0.02	-0.00
Retired	0.40	0.44	-0.08
Not in labour force	0.36	0.17	0.40
<i>Food security: Yes</i>	0.87	0.92	-0.17
<i>Veteran status: Yes</i>	0.19	0.27	-0.21
<i>Health insurance</i>			
Uninsured	0.06	0.06	0.01
Any private insurance	0.61	0.72	-0.23
Public insurance only	0.33	0.22	0.23
<i>Active cancer: Yes</i>	0.03	0.02	0.07
<i>Diabetes: Yes</i>	0.20	0.13	0.16
<i>Lung disease: Yes</i>	0.16	0.08	0.22
<i>Angina: Yes</i>	0.16	0.06	0.29
<i>Stroke: Yes</i>	0.09	0.05	0.12
<i>BMI category</i>			
Underweight (BMI \leq 18.5)	0.03	0.02	0.07
Normal weight (18.5 \leq BMI < 25)	0.31	0.38	-0.14

Variable	Averages for imputed datasets (before propensity score methods applied) (n = 19,971)		
	Pain (Exposed) (n = 4,073)	No Pain (Unexposed) (n = 15,898)	Standardized Mean Differences
Overweight ($25 \leq \text{BMI} < 30$)	0.35	0.40	-0.10
Obese 1 ($30 \leq \text{BMI} < 35$)	0.19	0.15	0.09
Obese 2 ($35 \leq \text{BMI} < 40$)	0.07	0.04	0.13
Obese 3 (BMI 40+)	0.04	0.01	0.14
<i>Heart condition: Yes</i>	0.30	0.18	0.26
<i>Arthritis: Yes</i>	0.81	0.47	0.87
<i>Smoker status</i>			
Never smoker	0.39	0.41	-0.03
Former smoker	0.42	0.43	-0.04
Current smoker	0.19	0.16	0.09
<i>Depressive symptoms (CESD score)</i>	2.75	1.34	0.61

Absolute standardized mean differences above the threshold of 0.1 (indicating imbalance between groups) are bolded. Pain (exposed) = moderate or severe pain; no pain (unexposed) = no or mild pain. N = 19,971; from the Health and Retirement Study, 1998.

BMI, body mass index.

The covariate balance plot in Figure 4.3 compares covariate balance after applying propensity score matching or IPW with covariate balance in the unadjusted data (covariate balance summaries for the alternative matching methods are also plotted in Fig. 4.3 and will be discussed later). The mean and range of ASMDs for each variable across the 20 imputations are plotted as points and lines respectively. The ASMDs for the unadjusted data (plotted in red) reflect the imbalance shown in Table 4.1, with some large values above 0.1. Good balance is achieved across all variables by both propensity score matching without replacement (plotted in green) and propensity score weighting (plotted in pink), as all green and pink points fall below the 0.1 threshold line. Therefore, all observed confounding appears to be adjusted for by the 2 propensity score methods under consideration.

Table 4.2 presents the pooled Cox proportional hazard models fitted to the unadjusted imputations before any propensity score methods were applied. In the unadjusted model with only pain exposure as a covariate, the HR for those with pain vs those without was 1.32 (95% CI: 1.26-1.38). Being exposed to pain was therefore associated with a significant and substantively large increase in mortality hazard compared with being unexposed. After adjusting for age, age squared, and sex, the estimated HR for pain exposure was 1.39 (95% CI: 1.33-1.46) (second model in Table 4.2). Thus, the significant association between pain exposure and mortality remained after adjusting for age and sex. After fitting the propensity score models and calculating propensity scores within each imputed dataset, on average 16.5 participants in the pain group and 49.2 in the no-pain group were outside the region of common support and were discarded before propensity score matching. Matching within each imputed dataset resulted in matched samples with an average size of 6,866.4 across the 20 imputed datasets, with an average of 4,056.2 participants in the pain group

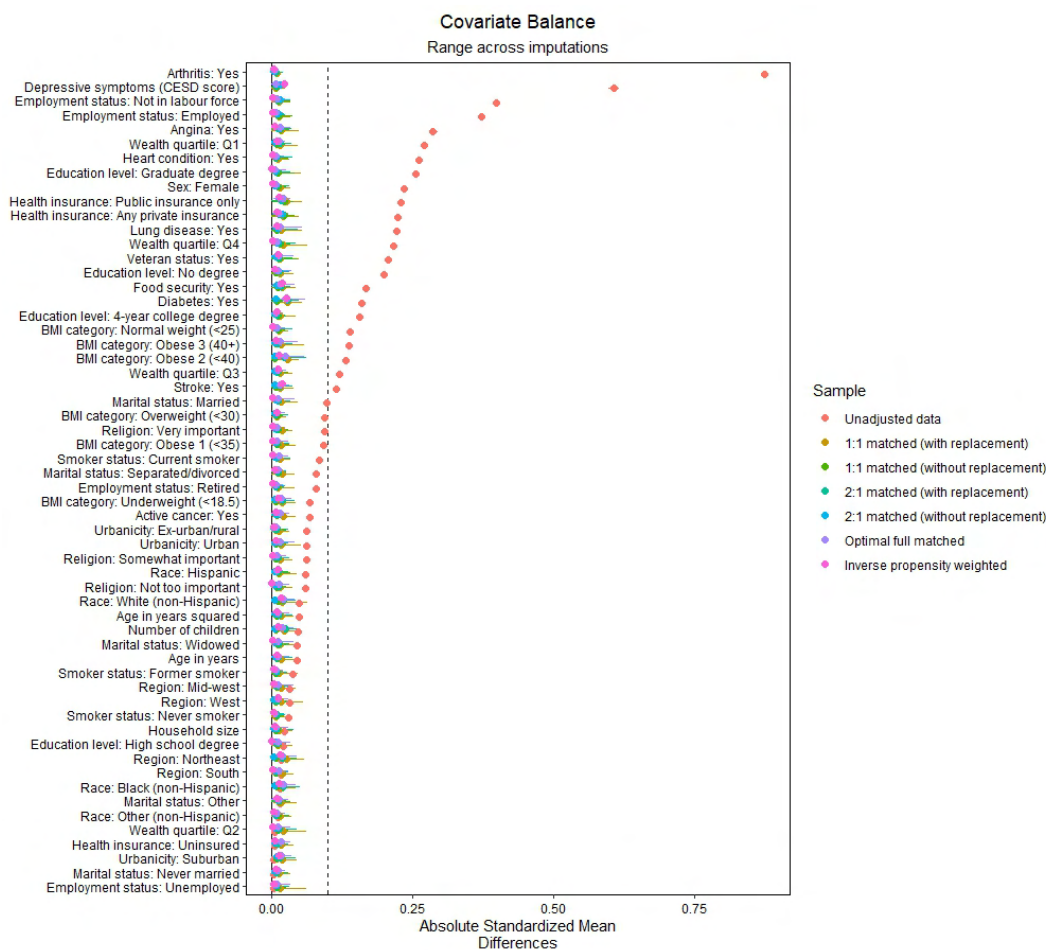


Figure 4.3: Covariate balance plot showing absolute standardized mean differences in covariate distributions between the pain (exposed) and no-pain (unexposed) groups across the 20 imputed datasets in the unadjusted data and after applying propensity score methods. The vertical dashed line at 0.1 shows the cut-off for values considered indicative of good balance. $N = 19,971$; from the Health and Retirement Study, 1998.

and 2,810.2 in the no-pain group. On average, less than one (0.4) participant in the pain group was unmatched and 13,038.7 participants in the no-pain group were unmatched.

Table 4.3 contains the pooled results of the Cox proportional hazards models fitted to the propensity score matched samples (when using the primary one-to-one matching without replacement method). The mortality HR for those with pain was 1.06 (95% CI: 0.99-1.14). While these results are

Table 4.2: Pooled hazard ratio effect estimates for pain exposure in 1998 on 20-year mortality from Cox proportional hazards models fitted to the imputed samples (no propensity score methods applied).

Analysis	Hazard ratio	95% confidence interval
Unadjusted	1.32	1.26-1.38
Adjusted for age, age squared, and sex	1.39	1.33-1.46

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018).

compatible with a modest positive effect of pain on mortality, they are also compatible with the possibility that pain exposure had no effect on mortality (i.e., the CI includes 1.0). It is worth noting that when compared with the 32% increase in mortality hazard estimated for the pain group in the unadjusted model in Table 4.2, the estimated increase in mortality hazard of 6% for the pain group in the matched sample represents an 81.25% reduction in estimated effect size. Adjusting for observed confounding using propensity score matching therefore appears to account for much and potentially all of the association between pain exposure and mortality found in the comparison models in Table 4.2. The Kaplan–Meier curves in Figure 4.4 reflect this finding, with the gap between the pain exposure and no-pain exposure curves for the unadjusted data appearing wider (Fig. 4.4(a)) than the gap between the averaged pain exposure and no-pain exposure curves for the matched datasets (Fig. 4.4(b)) visually showing that much or all of the mortality gap appears to be explained by the included confounders.

Inverse probability weighting resulted in an average effective sample of 4,073 participants in the pain group and 4,796.8 in the no-pain group across the 20 imputations. The results of the pooled Cox models fitted to the IPW samples, also shown in Table 4.3, were similar to the propensity score matching analysis, yielding an HR of 1.05 (95% CI: 0.99-1.10). Thus, the IPW analysis

Table 4.3: Pooled hazard ratio effect estimates for pain exposure in 1998 on 20-year mortality from Cox proportional hazards models, fitted to samples created using the primary propensity score matching technique and inverse probability weighting.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching without replacement	1.06	0.99-1.14
Inverse probability weighting	1.05	0.99-1.10
Regression-adjusted one-to-one matching without replacement	1.09	1.02-1.16
Doubly robust inverse probability weighted	1.06	1.00-1.12

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

is also suggestive of a modest causal effect of pain on mortality, although the 95% CI is again compatible with the possibility of no causal effect (contains 1.0).

The results of the regression-adjusted propensity score matching and the doubly robust IPW analyses are also summarized in Table 4.3. The HR for the regression-adjusted matching was 1.09 (95% CI 5 [1.02-1.16]). This analysis is more strongly suggestive of a causal effect of pain exposure on mortality, with the mortality hazard at any time over follow-up estimated to be 2% to 16% higher for those exposed to pain compared to if they had not been exposed to pain. Similarly, the doubly robust IPW method is most compatible with a modest causal effect of pain on mortality, with its 95% CI just containing 1.0 (HR = 1.06, 95% CI: 1.00-1.12).

The results from multiple alternative matching methods are provided in Table 4.4. Good balance was achieved by all additional matching methods, as shown in the covariate balance plot in Figure 4.3. In brief, almost all additional matching methods gave similar results to the one-to-one nearest neighbour matching without replacement shown in Table 4.3. All 95% CIs

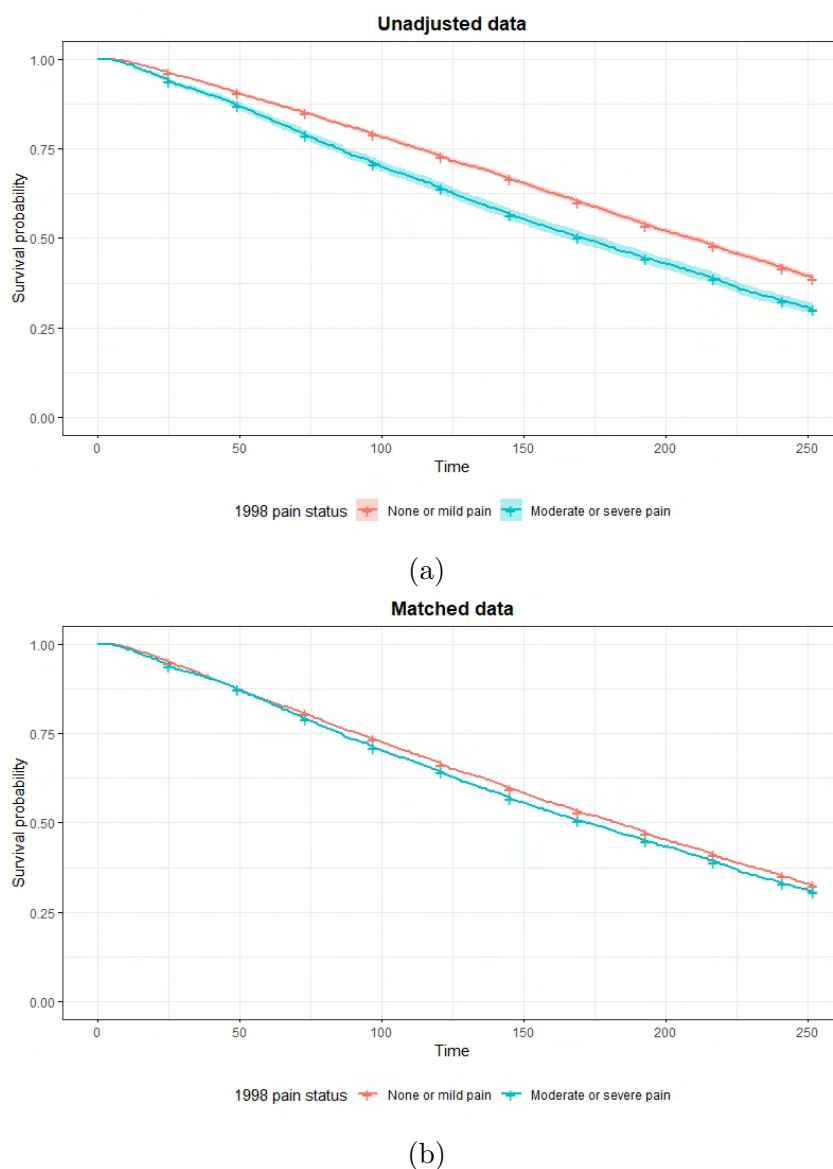


Figure 4.4: Kaplan–Meier survival curve plots stratified by pain exposure for (a) the unadjusted dataset, and (b) the 20 imputed and one-to-one propensity score matched without replacement samples. As neither pain exposure nor survival was imputed, the unadjusted data Kaplan–Meier curves were fitted using just the unadjusted dataset ($n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). For the one-to-one matched without replacement data Kaplan–Meier curves, mean survival probabilities were pooled across the 20 imputed and matched datasets and plotted. To our knowledge, there is no accepted convention for pooling variances across imputed and matched datasets for Kaplan–Meier plots, so confidence intervals are not calculated for this plot.

for the HR were compatible with a null or modest positive effect of pain on mortality hazard, with the exception of regression-adjusted 2-to-1 matching without replacement (HR = 1.08, 95% CI: 1.02-1.15), which resembled the results from the regression-adjusted one-to-one matching without replacement.

Table 4.4: Pooled hazard ratio effect estimates for pain exposure in 1998 on 20-year mortality from Cox proportional hazards models, fitted to matched samples created using alternative propensity score matching techniques.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.05	0.97-1.14
Two-to-one matching with replacement	1.05	0.97-1.13
Two-to-one matching without replacement	1.06	1.00-1.12
Optimal full matching	1.05	0.98-1.13
Regression-adjusted one-to-one matching with replacement	1.07	0.98-1.17
Regression-adjusted two-to-one matching with replacement	1.07	0.99-1.15
Regression-adjusted two-to-one matching without replacement	1.08	1.02-1.15
Regression-adjusted optimal full matching	1.06	0.97-1.14

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes vary depending on matching for each imputed sample.

Table 4.5 shows the results from analyses like the above except without adjusting for depressive symptoms. With depressive symptoms now positioned as a mediator rather than a confounder of the pain–mortality relationship, nearly all methods were most compatible with a modest positive causal effect of pain exposure on mortality, with HRs ranging from 1.06 to 1.12 and CIs nearly always excluding 1.0.

Overall, our sensitivity analyses indicate that our main results are robust to alternate modelling decisions. Repeating the analyses with HRS sample weights gave very similar results across all matching and IPW approaches (Supplementary Table A4.1). The results from analyses excluding arthritis

Table 4.5: Pooled hazard ratio effect estimates for pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques, without depressive symptoms included as a covariate in the propensity score models or Cox proportional hazards models.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.06	0.98-1.15
One-to-one matching without replacement	1.08	1.01-1.15
Two-to-one matching with replacement	1.06	0.99-1.14
Two-to-one matching without replacement	1.08	1.02-1.14
Optimal full matching	1.06	1.00-1.13
Inverse probability weighting	1.06	1.01-1.11
Regression-adjusted one-to-one matching with replacement	1.10	1.01-1.19
Regression-adjusted one-to-one matching without replacement	1.12	1.05-1.20
Regression-adjusted two-to-one matching with replacement	1.10	1.02-1.18
Regression-adjusted two-to-one matching without replacement	1.12	1.05-1.18
Regression-adjusted optimal full matching	1.08	1.01-1.16
Doubly robust inverse probability weighted	1.09	1.03-1.15

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

as a confounder were also very similar to the main analyses, except that the regression-adjusted matching methods without replacement also yielded 95% CIs containing 1.0 (Supplementary Table A4.2). Analyses of the effect of severe pain vs no, mild, or moderate pain (Supplementary Table A4.3), mild, moderate, or severe pain vs no pain (Supplementary Table A4.4), and severe pain vs no pain (Supplementary Table A4.5) again gave similar results to the main analysis (with pain exposure defined as moderate or severe pain). Reducing the follow-up length to 1, 5, or 10 years resulted in wider HR CIs (Supplementary Table A4.6) compared with those in Tables 4.2 and 4.3, likely due to fewer deaths occurring over follow-up. All estimated HR CIs

for the 1- and 5-year follow-ups contained 1.0, providing little insight into the strength or direction of the potential causal effect. The results for the 10-year follow-up were similar to the main 20-year follow-up analysis. Models that defined exposure as having moderate/severe pain and arthritis (and thus were particularly likely to capture chronic pain) yielded results very similar to our main analysis.

4.5 Discussion

This study used propensity score methods to rigorously explore how experiencing moderate or severe pain in 1998 influenced 20-year mortality in American older adults. Although we aimed to answer whether pain causally increases mortality risk, findings were equivocal. Models consistently yielded estimated HRs slightly above 1 and were therefore compatible with pain causing a small increase in mortality hazard, even after using propensity score methods to adjust for potential confounding by a large set of sociodemographic and health-related variables. Simultaneously, many models were also compatible with no causal effect, with only a minority of CIs excluding 1. On balance, our results are likely consistent with a modest causal effect of pain on mortality. However, replicating this finding using alternative data sources will be an important task for future research.

Before applying propensity score methods, we estimated Cox proportional hazard models of mortality including only pain as a predictor or including pain and adjusting only for age, age squared, and sex. Both models showed strong associations between pain and mortality, with mortality hazards over 30% higher for individuals with pain. However, in our many propensity score matching and inverse-probability weight models, much smaller associations

were found: mortality hazard was estimated to increase by 5% to 9% for individuals with pain, and many CIs were also compatible with no causal effect. For context, the causal effect we estimate is of a similar magnitude to the estimated effect of a 1-to-2-unit increase in BMI on mortality risk in the UK Biobank (HR per 1 unit increase in BMI: 1.03, 95% CI: 0.99-1.07) (Wade et al., 2018). Models that excluded depressive symptoms as a potential confounder were more consistently supportive of a modest causal effect (HRs between 1.06 and 1.12), signalling the sensitivity of our findings to this particular modelling decision. Overall, this study provides some evidence that pain itself, rather than the social or medical conditions that cause it, raises mortality risk. However, our findings are consistent with much or even perhaps all of the association being driven by upstream factors that increase the risk of both pain and mortality.

While existing research on pain and mortality primarily discusses association rather than causation, a handful of studies have explicitly examined the causal nature of the relationship. Some studies have found stronger evidence than ours to suggest that pain increases mortality risk. Evidence of a causal effect was found in the ELSA (a sister study of the HRS) (Smith et al., 2018), and a significant causal effect was found in a different American cohort including younger adults (aged ≥ 20 years) (Inoue et al., 2022). Our results may differ for a number of reasons, including different follow-up lengths and analytical approaches. In addition, we note that our findings may not necessarily generalize to other countries or age groups. For example, features of national healthcare systems (e.g., medical costs, or percent uninsured) may shape the pain-mortality link. Previous work has also found that pain correlates differ across age groups (Zajacova et al., 2021). The question of generalizability begs for further analyses using large international cohorts.

Moreover, previous studies used different confounder adjustment sets. For example, the ELSA study used just age, sex, education, and wealth as potential confounders and posed lifestyle, health, social, and psychological factors as potential mediators. In our study, the position of each variable in the underlying causal structure was carefully considered when identifying confounders. We aimed to remove confounder bias by adjusting for all measured confounders (Hernán and Robins, 2020), while avoiding the introduction of bias by adjusting for non-confounders (Van Zwieten et al., 2022; VanderWeele, 2009b). We used a comprehensive set of potential confounders and conducted sensitivity analyses when a variable’s position in the causal model was ambiguous. Our study also differs methodologically from earlier studies. Propensity score methods are preferable to traditional regression adjustments typically used in previous analyses, as they separate the study design from the outcome model by adjusting for confounding using propensity scores that are agnostic to the outcome, thus more closely emulating a randomised control trial (Amoah et al., 2020).

Our study has several implications for clinicians and public health advocates. Clinicians should be aware that chronic pain is predictive of mortality, suggesting that patients with pain should be closely monitored. Since the pain–mortality association is likely in part causal, optimally managing pain might improve not only patients’ quality of life but also their quantity of life. However, since our analyses also showed that much of the association between chronic pain and mortality is attributable to confounding, upstream factors that may raise the risk of both pain and death should also be addressed. Body mass index and depressive symptoms are 2 such factors adjusted for in our analyses. There is much research linking BMI to pain (Qian et al., 2021) and excess mortality (Flegal et al., 2013; Tobias and Hu, 2018). Previous

work has also suggested an association between depression and pain (Kroenke et al., 2011) and a dose-response effect between depressive symptoms and mortality (Schoevers et al., 2009; White et al., 2016). We also adjusted for several measures of low SES, which a voluminous literature has shown raises the risk of pain and mortality (Elo, 2009). These upstream factors may be more important drivers of excess mortality than pain itself and should be the subject of further investigation and public policies. Policies that reduce the risk of pain will, in many cases, also be policies to increase life expectancy.

One limitation of our study is that our pain variable, which specifies no particular duration, is not equivalent to the common definition of chronic pain as pain lasting over 3 months. We also lack information about the specific location or cause of pain. However, somewhat reassuringly, we note that the largest imbalances between our pain and no-pain groups before confounder adjustment were on characteristics known to be associated with chronic pain development, such as arthritis (Neogi, 2013; Walsh and McWilliams, 2014), education level, income level, and employment status (Mills et al., 2019). Future research may explore different datasets to better understand how findings vary across pain measures.

Our study is also limited by only considering pain at baseline. Research modelling pain exposure and confounders at multiple time points could clarify the potential cumulative causal effect of persistent pain on mortality risk. However, such modelling is complicated by exposure–confounder feedback loops, whereby historic pain is likely to causally influence confounders of pain (e.g., depressive symptoms) in the future. In such cases, regression models and propensity score methods are unable to adjust for confounding without introducing other biases (Hernán and Robins, 2020). Future work will require advanced causal analysis methods, such as the g-formula, to appropriately

handle feedback loops (Naimi et al., 2017).

Another limitation of all causal analyses of observational data is the unverifiable assumption that the causal structure underlying the analysis is correct and all confounders of the exposure-outcome relationship are appropriately adjusted for. We conducted sensitivity analyses to investigate how altering our confounder adjustment set would change results. While removing arthritis did not significantly alter interpretations, removing depressive symptoms shifted the results to more strongly suggest a positive effect of pain on mortality. This may be because depressive symptoms confound the pain–mortality relationship, so not adjusting for them is an error that creates a spurious association. However, it is also possible that depressive symptoms are a mediator of the pain–mortality relationship, so not adjusting for them “unblocks” the causal path from pain to mortality and permits correct estimation of the causal effect. As pain and some potential confounders were reported at the same wave in our data, it was difficult to determine temporal order. This could be addressed in future work using complex time-varying analyses which allow different factors to be posed as both confounders and mediators depending on temporal order. The lack of data on some potential confounders in the 1998 HRS Wave is also a limitation. For example, some psychological factors believed to affect pain outcomes (Crombez et al., 2023) were not measured.

Conducting mediation analyses (VanderWeele and Vansteelandt, 2014; Vansteelandt and Daniel, 2017) to investigate potential mediators of the pain–mortality relationship is another important aim for future work. Numerous mechanisms or pathways through which chronic pain may increase mortality risk can be postulated for further investigation. These may include biological pathways (e.g., pain potentially causing cardiovascular damage

(Fayaz et al., 2016)), pain management strategies linked to increased mortality risk (e.g., opioid analgesic use (Inoue et al., 2022)), or pain having a negative effect on overall health and well-being (e.g., by limiting physical activity (Chen et al., 2021)).

To our knowledge, this is the first study to use propensity score matching and IPW to investigate the potentially causal association between pain and mortality. We aimed to conduct a precise and comprehensive analysis of this relationship, adjusting for many reasonable confounders and applying various supplementary analyses to test the results' robustness. Our findings were mixed, comprising some analyses that clearly indicated a modest causal effect of pain on mortality, but also a greater number of analyses pointing in the same direction but compatible with no causal effect. This topic warrants further investigation with alternate data sources and modelling strategies. Nonetheless, the substantial attenuation of the observed association after confounder adjustment highlights the large role of potentially modifiable upstream risk factors for both pain and mortality. This work provides a basis for future studies examining the potential causal effect of pain on mortality in different countries and contexts, using different measures of pain, exploring potential mediators of the pain–mortality causal relationship, and considering the potential cumulative causal effect of pain at multiple time points.

4.6 Acknowledgements

This work analyses data from the HRS, which is conducted by the University of Michigan with financial support from the NIA (grant number NIA U01AG009740). Researchers seeking to access the HRS data may apply for access through the online HRS data portal: <https://hrsdata.isr.umich>.

edu/. The authors gratefully acknowledge the participants in the HRS study and all members of the HRS research team. The authors also wish to thank Professor Daniel Powers for helpful discussions and guidance regarding this study. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049 and of the National Institute on Aging of the National Institutes of Health under Award Number R01AG065351 (PI: H.G.-P.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4.7 Appendix 1: Propensity score methods

Propensity score matching

Propensity score matching involves pairing participants in the exposed group with participants in the unexposed group such that paired participants have close or identical propensity scores. This creates a matched sample of participants with similar propensities for pain, but with only one participant in a matched pair exposed to pain. One-to-one nearest neighbour caliper matching without replacement was used. This matching algorithm iterates through each pain-exposed participant and matches them to the unexposed participant with the most similar propensity score within a specified caliper distance. If there are no unexposed participants within the allowed caliper distance from the exposed participant's propensity score, the exposed participant remains unmatched. We applied a conservative caliper of 0.1 times the standard deviation of the propensity scores. When matching without replacement, once an unexposed participant is matched to an exposed participant they cannot be matched with another exposed participant. Nearest neighbour caliper matching without replacement has been found to outperform other matching algorithms (Austin, 2014a) including matching with replacement, which allows unexposed participants to be matched to multiple exposed participants. The region of common support refers to the range of propensity score values for which the exposed and unexposed groups overlap. A large overlap is required for successful matching. If a small number of scores fall outside this region it is recommended to discard them (Caliendo and Kopeinig, 2008; Dennison, 2019).

Alternative propensity score matching methods

The first alternative propensity score matching method applied was matching with replacement rather than without, meaning one unexposed participant could be matched to multiple exposed participants. We also applied two-to-one matching, meaning two unexposed participants were matched to each exposed participant. Both of these alterations can increase the number of matches made thus increasing the matched sample size. Optimal full matching was also conducted which involves subdividing the sample into a number of matched sets such that each set contains one exposed participant and ≥ 1 unexposed participant, or one unexposed participant and ≥ 1 exposed participant. The algorithm is considered optimal as the total distance in propensity scores between members of the same subsets is minimized (Hansen, 2004).

Inverse probability weighting

Inverse probability weighting (IPW) involves creating weights from the calculated propensity scores such that, after applying the weights, the exposed and unexposed groups have similar background characteristic distributions and differ only by exposure. As was also the case for propensity score matching, the causal effect we wish to estimate using IPW is the average treatment effect in the treated (ATT), meaning the average effect of pain exposure on mortality for those who were actually exposed to pain (Austin, 2014b). As those who were exposed to pain are the target population, these participants are assigned a weight of one. To weigh the unexposed group such that they have a distribution of background characteristics similar to the exposed group,

those who were not exposed to pain are assigned a weight using the formula

$$w = \frac{P(Z = 1|\mathbf{X})}{1 - P(Z = 1|\mathbf{X})}, \quad (4.1)$$

where $P(Z = 1|\mathbf{X})$ is the propensity score, the probability of exposure to pain ($Z = 1$) given the vector of observed background characteristics \mathbf{X} . If an unexposed participant has a very high propensity score this can result in a very large weight, which can increase the variability of the causal effect estimate. We thus stabilized the weights by multiplying each participant's weight by the proportion of participants in their exposure group (Austin and Stuart, 2015; Hernán and Robins, 2020).

Doubly robust models and regression-adjusted matching

While the accuracy of effect estimates obtained using propensity score matching or IPW alone rely on correct specification of the propensity score model for the exposure, doubly robust methods and regression-adjusted matching also adjust for confounding in the outcome model to safeguard against misspecification of the propensity score model (Kreif et al., 2013). In brief, doubly robust methods involve first fitting a propensity score model for the exposure to estimate inverse probability weights, then after applying the weights a second model is fitted for the outcome (in this case, a Cox regression model) with covariate adjustment for both confounders and exposure status. Only one of the two models need to be specified correctly for the effect estimate to be unbiased (Funk et al., 2011). Similarly, regression-adjusted matching involves first creating a propensity score matched sample as described above, then fitting an outcome model with covariate adjustment for confounders and exposure status (Austin, 2017).

4.8 Appendix 2: Sensitivity analyses

Table A4.1: Pooled hazard ratio effect estimates for pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques, *with HRS sample weights*.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.07	0.97-1.18
One-to-one matching without replacement	1.08	1.00-1.16
Two-to-one matching with replacement	1.06	0.98-1.15
Two-to-one matching without replacement	1.07	1.00-1.14
Optimal full matching	1.06	0.97-1.15
Inverse probability weighting	1.06	1.00-1.11
Regression-adjusted one-to-one matching with replacement	1.10	0.99-1.22
Regression-adjusted one-to-one matching without replacement	1.10	1.02-1.19
Regression-adjusted two-to-one matching with replacement	1.09	1.00-1.18
Regression-adjusted two-to-one matching without replacement	1.09	1.02-1.16
Regression-adjusted optimal full matching	1.07	0.97-1.17
Doubly robust inverse probability weighted	1.06	0.99-1.12

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

Table A4.2: Pooled hazard ratio effect estimates for pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques, *without arthritis* included as a covariate in the propensity score models or Cox proportional hazards models.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.03	0.95-1.11
One-to-one matching without replacement	1.04	0.97-1.11
Two-to-one matching with replacement	1.03	0.96-1.09
Two-to-one matching without replacement	1.04	0.99-1.10
Optimal full matching	1.03	0.96-1.10
Inverse probability weighting	1.03	0.98-1.08
Regression-adjusted one-to-one matching with replacement	1.05	0.96-1.14
Regression-adjusted one-to-one matching without replacement	1.07	1.00-1.14
Regression-adjusted two-to-one matching with replacement	1.04	0.97-1.12
Regression-adjusted two-to-one matching without replacement	1.06	1.00-1.12
Regression-adjusted optimal full matching	1.03	0.96-1.12
Doubly robust inverse probability weighted	1.04	0.98-1.10

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each n = 19,971; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

Table A4.3: Pooled hazard ratio effect estimates for *severe* pain exposure versus *no severe* pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques considered.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.07	0.93-1.22
One-to-one matching without replacement	1.07	0.94-1.21
Two-to-one matching with replacement	1.07	0.96-1.19
Two-to-one matching without replacement	1.07	0.97-1.18
Optimal full matching	1.07	0.97-1.18
Inverse probability weighting	1.09	1.00-1.18
Regression-adjusted one-to-one matching with replacement	1.13	0.97-1.30
Regression-adjusted one-to-one matching without replacement	1.12	0.98-1.27
Regression-adjusted two-to-one matching with replacement	1.12	1.00-1.26
Regression-adjusted two-to-one matching without replacement	1.12	1.01-1.25
Regression-adjusted optimal full matching	1.09	0.99-1.21
Doubly robust inverse probability weighted	1.12	1.03-1.22

Severe pain exposure = severe pain; no-severe pain exposure = no, mild, or moderate pain.

Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

Table A4.4: Pooled hazard ratio effect estimates for *any* pain exposure versus *no* pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques considered.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.05	0.97-1.14
One-to-one matching without replacement	1.07	1.02-1.14
Two-to-one matching with replacement	1.05	0.99-1.12
Two-to-one matching without replacement	1.08	1.02-1.13
Optimal full matching	1.05	0.98-1.12
Inverse probability weighting	1.04	0.99-1.10
Regression-adjusted one-to-one matching with replacement	1.08	0.99-1.16
Regression-adjusted one-to-one matching without replacement	1.09	1.03-1.16
Regression-adjusted two-to-one matching with replacement	1.07	1.00-1.15
Regression-adjusted two-to-one matching without replacement	1.09	1.03-1.15
Regression-adjusted optimal full matching	1.06	0.99-1.13
Doubly robust inverse probability weighted	1.07	1.01-1.13

Any pain exposure = mild, moderate or severe pain; no-pain exposure = no pain. Pooled results from 20 fully imputed datasets (each n = 19,971; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

Table A4.5: Pooled hazard ratio effect estimates for *severe* pain exposure versus *no* pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques considered.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.06	0.91-1.24
One-to-one matching without replacement	1.07	0.93-1.23
Two-to-one matching with replacement	1.06	0.94-1.20
Two-to-one matching without replacement	1.07	0.96-1.19
Optimal full matching	1.06	0.94-1.20
Inverse probability weighting	1.06	0.97-1.17
Regression-adjusted one-to-one matching with replacement	1.12	0.96-1.32
Regression-adjusted one-to-one matching without replacement	1.14	1.00-1.30
Regression-adjusted two-to-one matching with replacement	1.11	0.98-1.27
Regression-adjusted two-to-one matching without replacement	1.13	1.01-1.26
Regression-adjusted optimal full matching	1.08	0.94-1.23
Doubly robust inverse probability weighted	1.09	0.98-1.21

Severe pain exposure = severe pain; no-pain exposure = no pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

Table A4.6: Pooled hazard ratio effect estimates (confidence intervals in brackets) for pain exposure in 1998 on mortality over 1, 5, and 10-year follow-ups, estimated using all matching/weighting techniques considered.

Analysis	1 year follow-up	5 year follow-up	10 year follow-up
One-to-one matching with replacement	1.17 (0.69, 1.99)	1.04 (0.89, 1.21)	1.08 (0.97, 1.20)
One-to-one matching without replacement	1.16 (0.78, 1.73)	1.03 (0.91, 1.17)	1.08 (0.99, 1.19)
Two-to-one matching with replacement	1.16 (0.74, 1.80)	1.03 (0.90, 1.18)	1.07 (0.97, 1.18)
Two-to-one matching without replacement	1.16 (0.82, 1.64)	1.02 (0.91, 1.14)	1.08 (1.00, 1.16)
Optimal full matching	1.15 (0.76, 1.73)	1.03 (0.91, 1.18)	1.07 (0.98, 1.18)
Inverse probability weighting	1.14 (0.81, 1.59)	1.03 (0.93, 1.15)	1.06 (0.99, 1.15)
Regression-adjusted one-to-one matching with replacement	1.14 (0.66, 1.99)	1.07 (0.90, 1.27)	1.11 (0.98, 1.26)
Regression-adjusted one-to-one matching without replacement	1.15 (0.76, 1.75)	1.07 (0.94, 1.22)	1.12 (1.03, 1.23)
Regression-adjusted two-to-one matching with replacement	1.14 (0.71, 1.83)	1.06 (0.92, 1.23)	1.1 (1.00, 1.22)
Regression-adjusted two-to-one matching without replacement	1.15 (0.80, 1.66)	1.05 (0.94, 1.18)	1.11 (1.03, 1.21)
Regression-adjusted optimal full matching	1.11 (0.71, 1.73)	1.05 (0.91, 1.21)	1.09 (0.98, 1.21)
Doubly robust inverse probability weighted	1.12 (0.79, 1.59)	1.06 (0.95, 1.19)	1.09 (1.01, 1.18)

Pain exposure = moderate or severe pain; no-pain exposure = no or mild pain. Pooled results from 20 fully imputed datasets (each $n = 19,971$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

Table A4.7: Pooled hazard ratio effect estimates for *pain AND arthritis* exposure versus *no* pain exposure in 1998 on 20-year mortality estimated using all matching/weighting techniques considered.

Analysis	Hazard ratio	95% confidence interval
One-to-one matching with replacement	1.02	0.93-1.12
One-to-one matching without replacement	1.04	0.96-1.12
Two-to-one matching with replacement	1.02	0.94-1.10
Two-to-one matching without replacement	1.04	0.97-1.10
Optimal full matching	1.02	0.94-1.10
Inverse probability weighting	1.02	0.96-1.09
Regression-adjusted one-to-one matching with replacement	1.05	0.96-1.15
Regression-adjusted one-to-one matching without replacement	1.07	0.99-1.15
Regression-adjusted two-to-one matching with replacement	1.05	0.97-1.14
Regression-adjusted two-to-one matching without replacement	1.06	1.00-1.14
Regression-adjusted optimal full matching	1.04	0.96-1.13
Doubly robust inverse probability weighted	1.04	0.97-1.11

Pain AND arthritis exposure = moderate or severe pain AND arthritis; no-pain exposure = no or mild pain (with or without arthritis). Pooled results from 20 fully imputed datasets (each $n = 19,190$; from the Health and Retirement Study, 1998, followed through 2018). Sample sizes (effective sample sizes) vary depending on matching (weighting) for each imputed sample.

5 Normed Fit Indices for Latent Class Analysis with Large Sample Sizes

5.1 Introduction

As highlighted in Chapter 3 and Chapter 4, modelling sociodemographic disparities in pain and investigating the impact of pain on health and mortality outcomes are important ongoing areas of research. However, the exploration of these research questions is complicated by difficulties in effectively measuring pain. While statistical analyses often rely on a single variable (such as self-reported pain intensity) to represent the pain experience of study participants, there is a danger of oversimplifying or misrepresenting this complex, multi-dimensional condition by reducing pain experience to a single variable. A possible solution to this lack of a single comprehensive pain variable is to construct or identify a more holistic measure of pain using a combination of measures capturing different aspects of the pain experience. One potential methodological approach for identifying such a measure is latent class analysis (LCA), which has previously been used to capture pain subgroups characterised by multiple different variables related to pain (Dunn et al., 2006; O'Neill et al., 2018; Stynes et al., 2018).

While LCA represents a promising solution to this issue, in practice it can be challenging to apply latent class methodologies to cohort studies due to their typically large sample sizes. This chapter aims to address a methodological issue commonly faced when applying LCA methods to large cohort studies. Similar to other clustering methods, a key challenge for the researcher when fitting LCA models is selecting an optimal number of clusters (classes) to capture the underlying structure in the data. However, particularly with large sample sizes, the penalized fit criteria and likelihood ratio-based significance testing methods traditionally used for model selection in LCA can fail to clearly indicate one optimal model, or suggest an impractically large number of classes (Nylund et al., 2007). To address this LCA model selection issue, analogues of existing fit indices used in the structural equation modelling (SEM) literature are proposed for use with LCA. Two simulation studies are conducted as a preliminary investigation of the potential advantages and limitations of these fit indices for LCA, with a focus on large sample sizes. Note that, to further examine the performance of the proposed fit indices while also attempting to identify a holistic pain measure, LCA will later be applied to establish pain experience classes using the HRS dataset in Chapter 6.

The remainder of this chapter is organised as follows. Section 5.2 provides an introduction to the LCA model and further details on existing LCA model selection approaches and challenges. Section 5.3 provides background on SEM and describes some commonly used fit indices in the SEM literature. Section 5.4 details how a selection of SEM fit indices are adapted for use with LCA in this work. Section 5.5 presents the study designs, results, and a discussion for each simulation study. Finally, the overall findings and conclusions for this work are discussed in Section 5.6.

Some of the results in this chapter are presented in the following technical report:

Ryan, E., Dziak, J. J., Purtill, H., & Bray, B. C. (2023). Can a Normed Fit Index Assist with Model Selection in Latent Class Analysis with Large Samples? A Preliminary Investigation. (*Technical report available on PsyArXiv: <https://doi.org/10.31234/osf.io/3qzvm>*).

The author of this thesis (ER) was responsible for the design, implementation and interpretation of both simulation studies. All authors critically revised the technical report and approved the final version.

5.2 Latent class analysis (LCA)

LCA is a model-based clustering method for categorical data commonly used in the social, behavioural and health sciences. Applications of LCA methodologies in pain research have included identifying subgroups of biopsychosocial risk for pain development (O’Neill et al., 2018) and identifying subgroups characterised by multiple pain variables (Dunn et al., 2006; O’Neill et al., 2020; Stynes et al., 2018). Note that the tools used to aid class enumeration did not unanimously select the same number of classes in any of these studies. A more detailed introduction to the LCA model and LCA parameter estimation is included in the following subsections.

5.2.1 LCA model

The LCA algorithm identifies clusters or “classes” of a latent variable using data on a number of observed categorical variables (“indicator variables”),

which are believed to be causally related to the underlying latent variable and together indirectly measure the latent variable (Collins and Lanza, 2010). From a model-based clustering perspective, the observed variables form a multinomial joint distribution and the “classes” of the latent variable are the clusters to be identified. As an example, a visualisation of the typical structure of an LCA model with four observed indicator variables is included in Figure 5.1. The oval represents the latent class variable which characterises the clusters in the data, the squares represent the observed data on the four indicator variables, and the circles represent individual measurement error for each indicator variable. The direction of the arrows in this diagram represent the causal flow between variables.

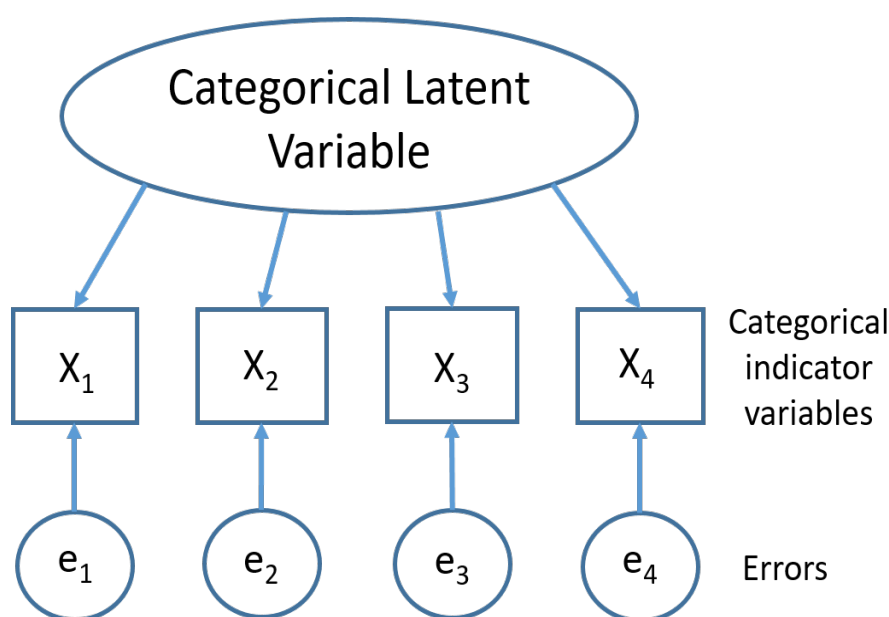


Figure 5.1: Example of a typical LCA model structure with four observed “indicator” variables. The directed arrows represent causal flow.

An important concept underlying LCA models is that the observed indicator variables are caused by the latent variable and measurement error, as depicted in Figure 5.1. The absence of arrows between the indicator

variables represents a fundamental assumption made by LCA models; *local independence*. Correlations between the indicator variables are assumed to be entirely explained by latent class membership, and so the indicator variables should be uncorrelated conditional on latent class membership. This means that within each latent class, the indicator variables are assumed to be independent (Visser and Depaoli, 2022). The local independence assumption has the advantage of simplifying the estimation of the LCA model, as discussed later in this section.

The parameters to be estimated when fitting an LCA model are the *latent class prevalences* and *item-response probabilities*. The former refers to the probability of belonging to each latent class, while the latter are the probabilities of observing the alternative responses to each indicator variable (or item) conditional on latent class membership. The vectors of latent class prevalences and item-response probabilities for an LCA model are typically represented by the Greek letters γ and ρ respectively.

To estimate these LCA parameters, a contingency table must first be created from the data. Consider a set of $j = 1, \dots, J$ observed categorical variables (the indicator variables). Let observed variable j have $r_j = 1, \dots, R_j$ possible response categories. By cross-tabulating these J observed variables, a contingency table with $W = \prod_{j=1}^J R_j$ cells is generated. Each cell of this contingency table represents a possible response pattern on the J observed variables, $\mathbf{y} = (y_1, \dots, y_J)$. Let \mathbf{Y} represent an array containing all possible response patterns. There is a probability associated with each response pattern \mathbf{y} , which will be represented by $P(\mathbf{Y} = \mathbf{y})$, such that $\sum P(\mathbf{Y} = \mathbf{y}) = 1$.

As the latent classes are mutually exclusive and exhaustive, for the latent class prevalences

$$\sum_{c=1}^C \gamma_c = 1,$$

where γ_c is the latent class prevalence associated with class c , and C is the total number of latent classes in the model. As each individual may only provide one response to each indicator variable j , it follows that

$$\sum_{r_j=1}^{R_j} \rho_{j,r_j|c} = 1,$$

where $\rho_{j,r_j|c}$ is the probability of observing response r_j on indicator variable j , conditional on membership in latent class c .

Recall that LCA models assume local independence between the indicator variables within each class, meaning that the indicator variables are independent of each other conditional on latent class membership. Thus, the probability of observing pattern \mathbf{y} in class c is given by the product

$$P(\mathbf{Y} = \mathbf{y} | L = c) = \prod_{j=1}^J \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)},$$

where $I(y_j = r_j)$ is an indicator function that equals 1 when the response to variable j is r_j , and equals 0 otherwise. Note that the latent class variable is represented by the letter L .

LCA models can thus be characterised by the following likelihood function

$$\prod_{i=1}^N P(\mathbf{Y} = \mathbf{y}_i | \rho, \gamma, C) = \prod_{i=1}^N \left[\sum_{c=1}^C \gamma_c \prod_{j=1}^J \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_{ij}=r_j)} \right],$$

where N denotes sample size. Maximum likelihood estimation is used to obtain estimates of the model parameters, typically using the expectation-maximisation (EM) algorithm (Dempster et al., 1977). In order to increase the chances of discovering the global maximum solution rather than a local

maximum, multiple sets of random initial values should be used when applying the EM algorithm. If the algorithm identifies the same solution for all or most of the sets of starting values, this adds confidence that the global maximum likelihood solution has been identified (Collins and Lanza, 2010).

5.2.2 LCA model selection

Determining how many clusters/classes best represent the underlying population structure, termed “class enumeration” in LCA, is an important and often challenging task. Typically, a decision must be made by the researcher as to what number of classes is most appropriate. A standard approach is to fit a range of candidate LCA models with different numbers of latent classes, and then apply a number of statistical tools to evaluate which model provides the closest and most parsimonious fit to the data.

One common set of tools used to aid the researcher with this task are information criteria such as the Akaike Information Criterion (AIC; Akaike, 1973) and Bayesian Information Criterion (BIC; Schwarz, 1978). Information criteria are relative measures of fit used to compare competing models in terms of how accurately and parsimoniously the models capture the covariance structure of the observed data (Collins and Lanza, 2010). The equations for the AIC and BIC are given by

$$AIC = G^2 + 2P \tag{5.1}$$

and

$$BIC = G^2 + [\log(N)]P \tag{5.2}$$

where G^2 is a likelihood ratio statistic calculated to assess how well the proposed latent class model fits the observed data, N is sample size, and

P is the number of model parameters estimated. Information criteria are sometimes referred to as penalised fit statistics, as a penalty based on the number of parameters P is applied to the goodness of fit measure G^2 . Multiple variations of these information criteria that apply different penalties have been developed, such as the sample size adjusted BIC (ABIC; Sclove, 1987).

Statistical significance tests including bootstrap likelihood ratio testing (BLRT; McLachlan and Peel, 2000) are also used for class enumeration. The distribution of the difference between G^2 likelihood ratio test statistics for an LCA model with k classes and a smaller model with $k - 1$ is undefined, thus bootstrapping is required to estimate the distribution. This is achieved by fitting models to many bootstrapped samples randomly generated from the smaller model with $k - 1$ classes. A p-value can then be obtained by determining where the observed G^2 difference falls in this distribution, with a p-value ≥ 0.05 suggesting that the larger k class model gives a superior fit (Collins and Lanza, 2010). The goal of both ICs and significance tests like the BLRT is to assess whether the proposed number of classes is sufficient to capture all of the statistically significant covariance in the data, or whether additional classes are required to properly characterise the underlying structure.

Class homogeneity and class separation are also desirable characteristics that can guide the selection of a final latent class model (Collins and Lanza, 2010). A latent class c is said to be highly homogeneous if one particular response pattern is very common among the members of class c , such that the response pattern could be said to be characteristic of c . In theory, perfect homogeneity occurs when all item response probabilities $\rho_{j,r_j|c} = 0$ or 1, and so only one distinct response pattern is observed in c . However, perfect homogeneity is unlikely to occur in real data. Class separation refers

to the degree to which the latent classes are distinct from one another. If a particular item response pattern has a high probability of occurrence in one class (and thus is characteristic of that class), that pattern will have a much lower probability of being observed in the other classes if the classes are well separated. Note that, while good latent class separation implies a high degree of class homogeneity, class homogeneity does not necessarily imply good class separation. Lack of homogeneity and poor class separation suggests that the indicator items used to estimate the model do not measure the latent variable of interest well (Masyn, 2013). Meanwhile, homogeneous classes and good class separation can aid interpretability, making it easier to assign meaningful and unique labels to the different classes.

Identifiability is also an important consideration when determining the number of classes for the final LCA model (Collins and Lanza, 2010). As mentioned in Subsection 5.2.1, the parameters of an LCA model are estimated by attempting to find the parameter values that maximise the likelihood function for the model using the EM algorithm. For the maximum likelihood (ML) solution to be identified, it is required that the degrees of freedom of the model be ≥ 1 . However, even with positive degrees of freedom the model may still be underidentified, meaning only some sets of starting values for the EM algorithm lead to the ML solution; or the model may be unidentified, meaning there is no unique ML solution. Underidentification occurs when there is a large amount of unknown information to be estimated (i.e., the model parameters) relative to the amount of known information (the available data). Identifiability issues are thus more likely to occur as the number of latent classes in the model, and thus the number of parameters to be estimated, are increased. On the other hand, the chances of identifying the ML solution are improved by having more known information. Large sample sizes, particularly

relative to the number of unique item response patterns for the indicator variables, are thus desirable. Indicator variables that are strongly related to the latent variable also provide more information. Thus, if the true underlying LCA model has good class homogeneity and class separation, identification issues are less likely.

Despite their prominence in latent class model selection procedures, information criteria and significance testing can prove unhelpful in some situations, particularly when fitting LCA models using a large sample size. The AIC specifically has been shown to perform poorly for large sample sizes (Nylund et al., 2007). When the dataset is large even an LCA model with practical significance (meaning a model which seems to fit reasonably well and make sense from the domain knowledge perspective) can have statistically significant lack of fit, due to small discrepancies between the observed and predicted response pattern contingency tables. This issue is common to many forms of null hypothesis testing, where a sufficiently large sample size can make small deviations from the ideal model “statistically significant” (Wasserstein and Lazar, 2016), even if the proposed model provides practically useful insights.

As a result, when sample size is large, the traditional methods for determining how many latent classes are sufficient may suggest an impractically large number of classes. This situation leads to a number of issues, as these methods aim to identify the absolutely “correct” LCA model rather than a model that fits reasonably well and is practically useful and interpretable. For example, a large number of classes can cause identification issues when estimating model parameters using maximum likelihood. Many small classes can also make it difficult to interpret the LCA model in the real-world context of the data.

5.3 Structural equation modelling (SEM)

Structural equation modelling (SEM) refers to a range of multivariate methods used to model the relationships between continuous observed and latent variables, such as confirmatory factor analysis and exploratory factor analysis. In brief, SEM consists of a measurement model concerned with how the latent variables or constructs are measured by observed variables, and a structural model to estimate the directional (possibly causal) relationships between the latent variables (Bowen and Guo, 2011). A detailed introduction to SEM methodology can be found in Bowen and Guo (2011).

In some ways, LCA with categorical data is analogous to structural equation modelling (SEM) approaches for normally distributed data (Sánchez et al., 2005). Both LCA and SEM with normally distributed data attempt to model an underlying latent variable using data on observed variables which are believed to be caused by the latent variable. It is thus proposed that model selection methods applied in SEM may also be helpful to guide model selection in LCA. Subsection 5.3.1 below presents a summary of fit indices commonly used to assess and compare model fit in SEM. The adaption of the discussed SEM fit indices for use with LCA is then proposed in Section 5.4.

5.3.1 SEM model selection

Multiple indices for measuring qualitative fit have been developed for SEM with normally distributed data (Hooper et al., 2008). These fit indices describe the degree to which a proposed model fits the data, without performing a formal test for statistically significant lack of fit. One of the oldest of these indices is the Bentler-Bonett normed fit index (NFI) (Bentler and Bonett, 1980). The NFI measures the lack of fit of the proposed model in comparison

to the lack of fit of a null model, on a scale from zero to one. The “lack of fit” is captured by a chi-squared test statistic for lack of fit. The formula for the NFI is:

$$NFI = \frac{\chi_{null}^2 - \chi_{candidate}^2}{\chi_{null}^2}. \quad (5.3)$$

The maximum NFI value of one will be achieved by a “saturated” model which has enough parameters to exactly reproduce the observed data patterns, but is unlikely to be interpretable or generalisable due to overfitting ($\chi^2 = 0$ for such a model). At the other end of the scale an NFI of zero will be calculated for the null model, which will have the highest possible χ^2 statistic value. The null model is typically the most simple model possible which assumes all variables are unconditionally independent. Thus, the NFI quantifies the level of model fit of a proposed model in comparison to these two extremes.

A variation of the NFI called the non-normed fit index (NNFI), or the Tucker-Lewis index (Tucker and Lewis, 1973), is also used in SEM. This index introduces a penalty for model complexity based on the degrees of freedom (df) of the candidate model. While this adaptation puts greater emphasis on parsimony, it is at a cost of no longer scaling between zero and one. The formula for the NNFI is given as

$$NNFI = \frac{\frac{\chi_{null}^2}{df_{null}} - \frac{\chi_{candidate}^2}{df_{candidate}}}{\frac{\chi_{null}^2}{df_{null}} - 1}. \quad (5.4)$$

Both 0.90 and 0.95 have been suggested as cutoffs for interpreting the NFI, with values above these cutoffs suggestive of “good” model fit (Hooper et al., 2008). Similarly, NNFI values above 0.95 are commonly interpreted as an indicator of “good” fit. Note that these informal “rule-of-thumb” guidelines for interpreting the values of fit indices in SEM have mostly been established using Monte Carlo simulation experiments (Hu and Bentler, 1999).

The NFI and NNFI are measures of relative fit, comparing the lack of fit of a candidate model with that of the null and saturated models. Multiple measures of absolute fit are also used in SEM. These measures assess how well the proposed model fits the sample data without comparison with a baseline model. One popular index is the root mean squared error of association (RMSEA), which adjusts for both model complexity and sample size (n). The RMSEA is calculated as follows:

$$RMSEA = \sqrt{\frac{\chi^2 - df}{(n - 1)df}}. \quad (5.5)$$

Informally, a RMSEA value below 0.07 is considered desirable (Hooper et al., 2008).

The standardized root mean squared residual (SRMR) is an alternative measure of absolute fit to the RMSEA also used in SEM. This index measures the discrepancy between the sample covariance matrix and the hypothesized covariance matrix for the candidate SEM model. The SRMR is calculated using the following formula:

$$SRMR = \sqrt{\left[2 \sum_{i=1}^p \sum_{j=1}^i \frac{s_{ij} - \hat{\sigma}_{ij}}{s_{ii}s_{jj}} \right]^2 / p(p+1)}, \quad (5.6)$$

where p = the number of observed variables; s_{ij} = the observed covariances; and $\hat{\sigma}_{ij}$ = the covariances reproduced by the proposed model (Hu and Bentler, 1999). SRMR values fall between zero and one, and values below 0.08 are generally accepted as indicating good fit (Hooper et al., 2008).

The NFI, NNFI, RMSEA, and SRMR all have the common benefit of describing model fit on a relative scale from poor to good, rather than making a dichotomous decision to reject or accept a model on the basis of statistical significance. It thus may be useful to have analogues to these indices for LCA modelling. Such tools may help to avoid the issue of practically useful LCA

models (with a reasonably small number of classes) being rejected on the basis of statistical significance due to potentially small discrepancies in model fit.

5.4 Adapting SEM fit indices for LCA

To adapt the fit indices used in SEM for LCA models an LCA equivalent to the χ^2 test statistic used in SEM must be identified. There are two chi-squared distributed LCA fit statistics which could potentially be used; the Pearson chi-squared statistic,

$$\omega^2 = \sum_{x=1}^{n_z} \frac{(f_z - \hat{f}_z)^2}{\hat{f}_z}, \quad (5.7)$$

and the deviance chi-squared statistic (mentioned earlier in Subsection 5.2.2),

$$G^2 = 2 \sum_{x=1}^{n_z} f_z \log \left(\frac{f_z}{\hat{f}_z} \right), \quad (5.8)$$

where n_z is the number of cells in the observed data contingency table, f_z is the observed proportion of individuals in cell z of the contingency table, and \hat{f}_z is the proportion of individuals in cell z as predicted by the fitted LCA model. The Pearson chi-squared statistic ω^2 is the square of Cohen's effect size for comparing the candidate model against a saturated alternative, while the deviance chi-squared statistic G^2 is the Kullback-Leibler distance from the alternative (Dziak et al., 2014). As both of these chi-squared fit statistics are asymptotically chi-squared under the null hypothesis, it is likely that they will give similar results when the sample size is large. The deviance G^2 will be used in this work as it is automatically generated by most LCA software (Lanza et al., 2007; Vermunt and Magidson, 2013).

The degrees of freedom of an LCA model depends on the number of cells in the data contingency table (n_z) and the number of parameters to be estimated by the model (P). In Collins and Lanza (2010), the degrees of freedom for an

LCA model are given by:

$$df = n_z - P - 1. \quad (5.9)$$

By substituting the deviance chi-squared statistic G^2 and degrees of freedom for LCA into the SEM formulas, analogues of the NFI, NNFI and RMSEA for LCA can be defined as follows:

$$NFI_{LCA} = \frac{G_{null}^2 - G_{candidate}^2}{G_{null}^2}, \quad (5.10)$$

$$NNFI_{LCA} = \frac{\frac{G_{null}^2}{df_{null}} - \frac{G_{candidate}^2}{df_{candidate}}}{\frac{G_{null}^2}{df_{null}} - 1}, \quad (5.11)$$

$$RMSEA_{LCA} = \sqrt{\frac{G^2 - df}{(n - 1)df}}. \quad (5.12)$$

As mentioned in the previous section, the SRMR measures the discrepancy between the sample covariance matrix and the covariance matrix produced by the candidate SEM model. An analogous LCA index might instead consider the discrepancy between the sample contingency table and the contingency table predicted by the fitted LCA model. While investigating the potential for adapting the SRMR index for use in LCA, an existing but uncommonly used LCA measure called the Dissimilarity Index was identified in the literature (DI; Vermunt and Magidson, 2013). The DI is also designed to measure the difference between observed and estimated contingency table cell frequencies, and is calculated using the following formula:

$$DI = \frac{(\sum_z^{n_z} |m_z - \hat{m}_z|) + (N - \sum_z^{n_z} \hat{m}_z)}{2N}, \quad (5.13)$$

where m_z is the observed count of individuals in cell z and \hat{m}_z is the predicted count of individuals in cell z . As any attempt to create an SRMR analogue for LCA would likely be very similar in principle to the existing DI, the adaptation of this fit index to LCA was not pursued further.

Versions of the NFI, NNFI, and RMSEA which are potentially analogous to their SEM counterparts have now been defined for use in LCA. However, an understanding of how these indices might perform and be interpreted when applied to real-world analyses must be developed. In the next section, the behaviour of the proposed fit indices under different sample sizes and LCA population structures is explored using two simulation studies.

5.5 Simulation studies

5.5.1 Simulation study 1

As a first step, a simulation study was conducted using a selection of simple, plausible underlying latent class structures as data generation models. The aim of this preliminary study was to investigate how the proposed LCA fit indices behaved when numerous factors defining the underlying model were varied, including class sizes, class separation, and class homogeneity. As the challenge of class enumeration for large sample sizes was of particular interest when developing these LCA fit indices, a wide range of sample sizes were also considered. For the various data generation conditions the performance of the proposed fit indices were compared with two commonly used existing measures, the AIC and BIC. Performance was measured by how often each measure correctly selected the fitted model with the same number of classes as the data generation model, rather than an under-specified or over-specified model, based on pre-specified selection criteria. This study provided some initial insights into how the proposed fit indices may behave and be utilised in practice. Findings from this initial simulation study also informed the design of the second simulation study (discussed later in Subsection 5.5.2).

5.5.1.1 Design

Population models with four latent classes, no covariates, and eight dichotomous indicator items were used to generate the samples. A broad range of sample sizes ($N = 500$; 1,000; 10,000; and 40,000) were used in the simulations, as the challenge of selecting an LCA model with an appropriate number of classes for large samples was the motivation for this study. To increase the generalisability of the results, various possible structures and patterns of item response probabilities were accounted for in the data generation process, as well as different class sizes.

Two different class-specific item response probability structures were used to generate the sample data ('Structure A' and 'Structure B'), as detailed in Table 5.1. For Structure A each class is distinguished by high probabilities of 'yes' responses on a different pair of items and low probabilities of 'yes' responses on the remaining items, representing good class separation. For Structure B classes are not as well separated, as high probabilities of a 'yes' response for certain items are characteristic of more than one class.

Class homogeneity was also varied using 'strong' and 'weak' item response patterns. In the 'strong' case, item response probabilities for a 'yes' response were set to 0.9 for high or 0.1 for low for each binary item, resulting in high class homogeneity. In the 'weak' case, item response probabilities were set to 0.7 or 0.3 respectively, corresponding to more heterogeneous responses for individuals in the same class. The class-specific probabilities of a 'yes' response to each dichotomous item are listed in Table 5.1. The possibility of class membership probabilities being equal (0.25 for each class) or unequal (0.4, 0.3, 0.2, and 0.1 for classes 1, 2, 3, and 4 respectively) were also considered.

This data generation procedure was designed to reflect some realistic real-world latent class structures and was similar to procedures used in previous

Table 5.1: Summary of the LCA data generation models used in simulation study 1. The class-specific probabilities of a ‘yes’ response for each item are detailed for both the strong and weak patterns (weak in brackets).

		Class			
		1	2	3	4
<i>Latent Class Membership Probability</i>	Equal classes	.25	.25	.25	.25
	Unequal classes	.40	.30	.20	.10
Structure	Indicator	<i>Item-Response Probability for ‘Yes’ Response</i>			
A	1	.9 (.7)	.1 (.3)	.1 (.3)	.1 (.3)
	2	.9 (.7)	.1 (.3)	.1 (.3)	.1 (.3)
	3	.1 (.3)	.9 (.7)	.1 (.3)	.1 (.3)
	4	.1 (.3)	.9 (.7)	.1 (.3)	.1 (.3)
	5	.1 (.3)	.1 (.3)	.9 (.7)	.1 (.3)
	6	.1 (.3)	.1 (.3)	.9 (.7)	.1 (.3)
	7	.1 (.3)	.1 (.3)	.1 (.3)	.9 (.7)
	8	.1 (.3)	.1 (.3)	.1 (.3)	.9 (.7)
B	1	.9 (.7)	.1 (.3)	.9 (.7)	.1 (.3)
	2	.9 (.7)	.1 (.3)	.9 (.7)	.1 (.3)
	3	.9 (.7)	.1 (.3)	.9 (.7)	.1 (.3)
	4	.9 (.7)	.1 (.3)	.9 (.7)	.1 (.3)
	5	.9 (.7)	.1 (.3)	.1 (.3)	.9 (.7)
	6	.9 (.7)	.1 (.3)	.1 (.3)	.9 (.7)
	7	.9 (.7)	.1 (.3)	.1 (.3)	.9 (.7)
	8	.9 (.7)	.1 (.3)	.1 (.3)	.9 (.7)

LCA simulation studies. Specifically, data generation models with four latent classes, eight binary indicator items, and varying levels of class separation and class homogeneity were also used in an earlier study of LCA class enumeration approaches (Nylund et al., 2007).

1,000 replication samples were generated for each data generation model and sample size combination. LCA models with the correct number of classes (four) were then fitted to each replication sample. 3-class and 5-class LCA models were also fitted to each sample to compare average fit index values for the true and misspecified LCA models. Neighbouring class models only (within one class of the correct number of classes) were focused on for this preliminary

investigation, as it was of interest to explore how well the proposed fit indices distinguished the correct model from neighbouring models for different sample sizes compared to the AIC and BIC. To aid in identifying the maximum likelihood solution, 100 sets of random parameter starting values were used to fit each model (Collins and Lanza, 2010).

The proposed NFI, NNFI, and RMSEA fit indices were computed for all fitted models in simulation study 1. The AIC and BIC were also calculated for each model for comparison. Means and 95% coverage intervals were calculated for each fit index across all replication samples for each model. The means and 95% coverage intervals for each measure were then plotted for each sample size. This provided a general way to visually examine each model selection tool across the true and misspecified latent class models.

To further investigate the behaviour of the proposed fit indices compared to the AIC and BIC, the number of classes selected in each replication sample by each model selection tool was recorded using several decision rules. First, the number of classes was selected based on the standard approach of selecting the model with the lowest value of AIC/BIC. Second, a slightly more parsimonious way of using the AIC and BIC was applied. Burnham and Anderson (2004) considered any model with an AIC within 2 units of the lowest observed AIC as being plausible, while Raftery (1995) considered a BIC difference of 2 units or less to be “weak” evidence for choosing one model versus another. Thus, based on these criteria, the number of classes for the most parsimonious model with an AIC/BIC that was no more than 2 units higher than the lowest AIC/BIC was selected. Note that because the AIC and BIC are based on log-likelihoods, differences between AICs or BICs have similar meanings regardless of the total size of the criterion; that is, differences are more meaningful than ratios. For the NFI and NNFI, the commonly suggested cutoff values used to identify

acceptable model fit in SEM were applied. The smallest number of classes for which $NFI > 0.90$ was selected for each set of models. Similarly, the model with the smallest number of classes for which $NNFI > 0.95$ was selected. A selection rule based on the RMSEA was not applied as this index proved impractical for use with LCA. Further details on the performance of the RMSEA are included in the Results section below.

Data generation and fit index calculations were carried out using R (R Core Team, 2022). All LCA models were fitted using LatentGOLD (Vermunt and Magidson, 2021) version 6.0.

5.5.1.2 Results

The simulation study quickly highlighted a limitation of the proposed RMSEA index for LCA. The index was found to be undefined for many of the LCA models fitted to the simulated datasets, caused by the degrees of freedom exceeding the value of the G^2 statistic for the model in the proposed RMSEA formula (Equation 5.12). The percentage of models for which the RMSEA was undefined for each set of simulations is summarised in Table 5.2. The issue became more prevalent as the number of classes in the models increased, with the RMSEA being undefined for up to 100% of the 4- and 5-class models fitted to some sets of simulations. Due to this severe limitation, it was concluded that the proposed RMSEA statistic would be unsuitable for LCA. The remainder of this Results section will thus focus on the behaviour of the NFI and NNFI in simulation study 1.

As described in the above study design, a number of important factors were varied in the LCA models used for data generation to increase the generalisability of the results. This gave eight unique data generation models [2 structures (A/B) \times 2 item response patterns (strong/weak) \times 2 class

sizes (equal/unequal)]. Varying some of these factors appeared to have little influence on the behaviour of the fit indices under consideration in this specific simulation study. In particular, the behaviours of the proposed NFI and NNFI were very similar regardless of whether the data generation model had equal or unequal class sizes. To avoid the repetitiveness of reporting multiple sets of similar results, only the summary figures and tables for the simulations with *equal* class sizes are included in the main text of this chapter. Similar summary figures and tables for the models with *unequal* class sizes are included in Appendix A for reference.

Table 5.2: Percentage of 3-, 4-, and 5-class models for which the RMSEA was undefined in each set of 1000 simulations.

Class sizes		Equal			Unequal		
Structure/ pattern	Sample size	Number of classes					
		3	4	5	3	4	5
Structure A/ Strong pattern	500	0.0%	100.0%	100.0%	0.2%	100.0%	100.0%
	1,000	0.0%	99.2%	99.9%	0.0%	99.9%	100.0%
	10,000	0.0%	44.2%	64.0%	0.0%	48.5%	68.2%
	40,000	0.0%	42.7%	59.9%	0.0%	40.6%	58.5%
Structure A/ Weak pattern	500	4.6%	13.7%	20.5%	8.7%	16.1%	25.0%
	1,000	2.0%	19.3%	30.1%	6.8%	20.7%	32.7%
	10,000	0.0%	46.3%	64.5%	0.0%	46.2%	63.0%
	40,000	0.0%	51.6%	70.5%	0.0%	52.2%	69.5%
Structure B/ Strong pattern	500	0.0%	100.0%	100.0%	0.0%	100.0%	100.0%
	1,000	0.0%	93.5%	98.5%	0.0%	93.8%	98.4%
	10,000	0.0%	24.7%	39.7%	0.0%	25.6%	40.3%
	40,000	0.0%	44.8%	60.4%	0.0%	41.8%	58.4%
Structure B/ Weak pattern	500	2.6%	7.2%	11.5%	3.3%	9.0%	13.6%
	1,000	4.8%	20.6%	31.7%	10.0%	23.3%	33.0%
	10,000	0.0%	47.9%	65.7%	0.0%	48.9%	65.4%
	40,000	0.0%	50.6%	68.4%	0.0%	48.8%	68.4%

The distributions of the AIC, BIC, NFI, and NNFI values calculated for the LCA models fitted to the structure A strong pattern simulations with equal class sizes are plotted in Figure 5.2. Means are plotted as circles and 95%

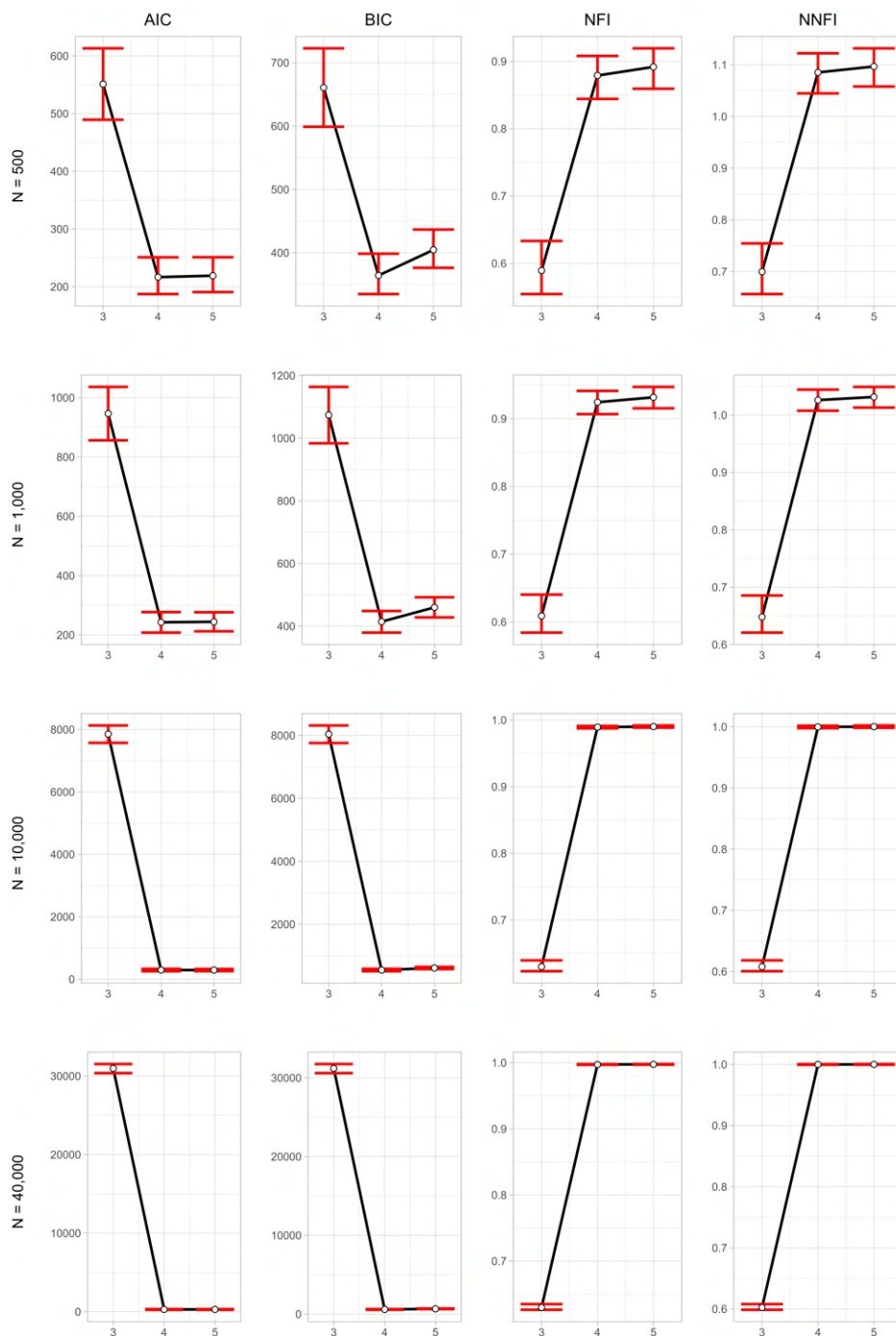


Figure 5.2: Distributions of fit indices values for the **structure A strong pattern** simulations with **equal class sizes**. *Notes:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

coverage intervals are plotted as red bars. The x-axes denote the number of classes (3-5) in the fitted models while the y-axes denote fit indices values. Each row of plots corresponds to a different sample size.

Regardless of sample size, all four indices appeared to clearly distinguish between the under-specified 3-class model and the correctly-specified 4-class model, with a visibly large jump in values indicating the superiority of the 4-class model (large decrease in AIC/BIC; large increase in NFI/NNFI). The change in mean values between the 4-class and 5-class models were much smaller for all four indices across all sample sizes, with some overlap in the 95% coverage intervals for the 4-class and 5-class models. This tapering off gave the appearance of an “elbow” in the plots at 4 classes.

The application of commonly used selection rules to the structure A strong pattern simulations in Table 5.3 gives further insight into the performance of each fit index for this set of simulations. A 3-class model was never chosen over a 4-class model by any fit measure criteria across the different sample sizes, reflecting the comparatively poor fit of the 3-class models evident in Figure 5.2. For the smallest sample size ($N = 500$), choosing a number of classes based on the $NFI \geq 0.9$ cutoff resulted in the over-specified 5-class model being favoured over the correctly specified 4-class model in 91% of simulations. However, this NFI criteria has essentially 100% accuracy in correctly selecting the 4-class models for sample sizes $\geq 1,000$. In comparison, the AIC selection rules displayed a tendency to sometimes favour the over-specified 5-class model across all sample sizes. This overestimation was more common when the model with the lowest AIC was selected, rather than the most parsimonious model with AIC within 2 units of the minimum AIC value. This tendency worsened as sample size increased, with the “lowest AIC” and “AIC within 2” selection rules choosing the 5-class model for 44.8% and 28.5% of simulated samples

respectively when $N = 40,000$.

Table 5.3: Percentage of **structure A strong pattern** simulated data sets with **equal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
$N=500$	Lowest AIC	0.0%	79.4%	20.6%
	AIC within 2	0.0%	88.9%	11.1%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	9.0%	91.0%
	NNFI >.95	0.0%	100.0%	0.0%
$N=1,000$	Lowest AIC	0.0%	69.2%	30.8%
	AIC within 2	0.0%	83.3%	16.7%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	99.6%	0.4%
	NNFI >.95	0.0%	100.0%	0.0%
$N=10,000$	Lowest AIC	0.0%	57.0%	43.0%
	AIC within 2	0.0%	73.2%	26.8%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%
$N=40,000$	Lowest AIC	0.0%	55.2%	44.8%
	AIC within 2	0.0%	71.5%	28.5%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

The NNFI and BIC criteria proved most adept at identifying the correct number of classes for this set of simulations. Choosing a number of classes based on the NNFI ≥ 0.95 selection criteria identified the correct 4-class solution for 100% of simulations across all sample sizes. Both BIC selection

criteria (choosing the model with the lowest BIC value; choosing the most parsimonious model with BIC value within 2 units of the minimum BIC value) also achieved 100% accuracy regardless of sample size. Overall, while the NFI and NNFI proved more reliable than the AIC for this specific data generation model across a range of sample sizes (except the smallest sample size for the NFI), the BIC criteria performed just as well.

Next, consider the simulations which also used the strong item response pattern and equal class sizes, but were generated using structure B rather than structure A. Observing the fit indices summary plots for the structure B strong pattern simulations in Figure 5.3, very similar “elbow” shaped plots which mirror the plots in Figure 5.2 are noted. Table 5.4, which contains a summary of how often 3-, 4-, or 5-class models were selected by applying the various fit index selection rules to the structure B strong pattern simulation models, also closely reassembles the earlier results for the structure A strong pattern simulations (Table 5.3). Again, the NNFI and both BIC selection criteria selected the correct 4-class models for all simulations, while the NFI also achieved 100% accuracy when sample size was $N \geq 1,000$. Both AIC criteria again tended to sometimes favour the over-specified 5-class models across all sample sizes. This overestimation occurred at a similar rate in the structure B strong pattern simulations as for the structure A strong pattern simulations.

Thus, when the item response pattern was “strong”, the performances of the existing and proposed fit indices appeared to be consistent for the two latent class model structures across sample sizes considered in this simulation study. The NFI, NNFI, and BIC all out-performed the AIC when interpreted using the specified selection rules. Next, how the various fit indices behaved when the item response pattern in the underlying LCA model was “weak”

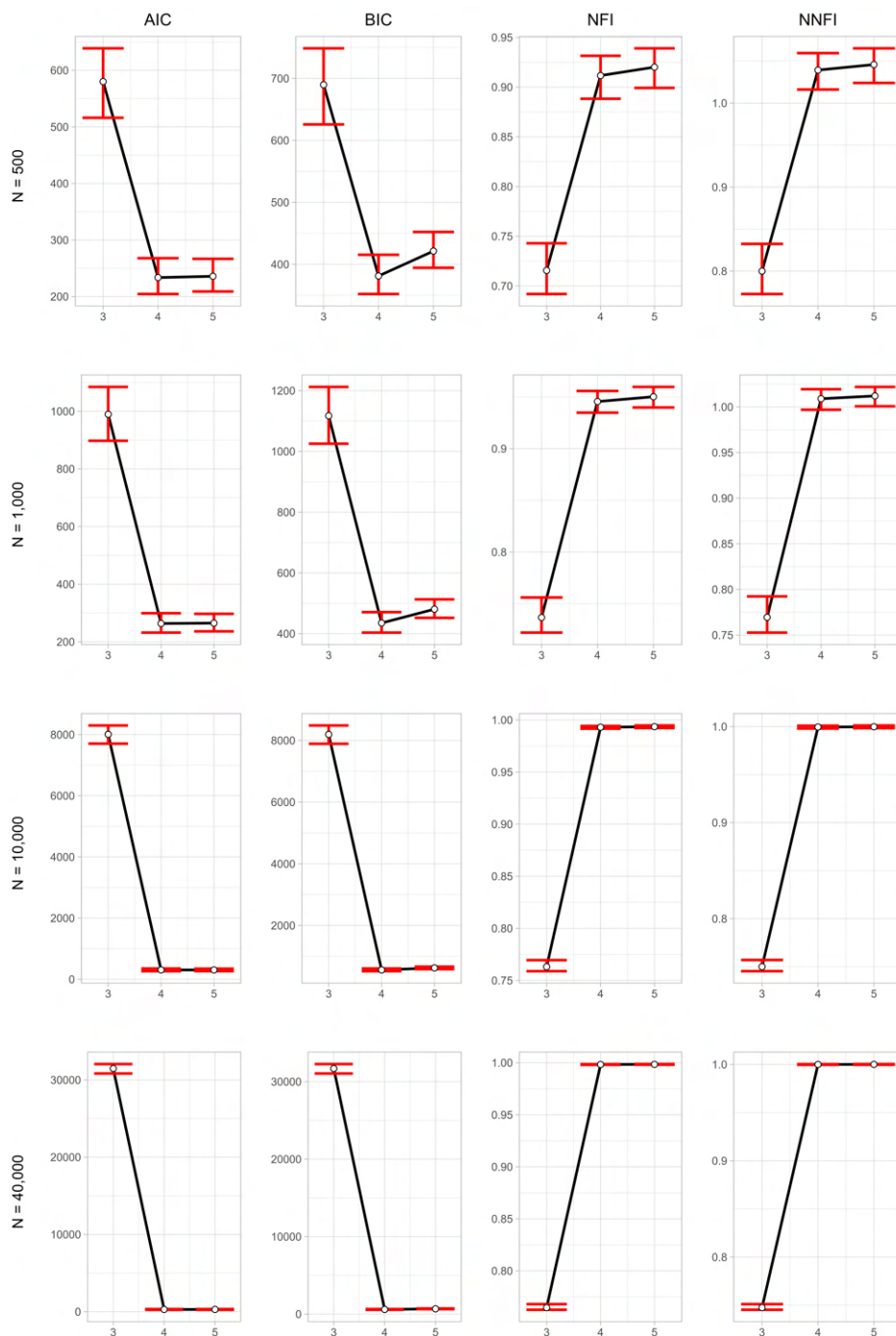


Figure 5.3: Distributions of fit indices values for the **structure B strong pattern** simulations with **equal class sizes**. *Notes:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

rather than “strong” is reported.

Table 5.4: Percentage of **structure B strong pattern** simulated data sets with **equal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	0.0%	78.2%	21.8%
	AIC within 2	0.0%	89.0%	11.0%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	86.0%	14.0%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=1,000</i>	Lowest AIC	0.0%	69.3%	30.7%
	AIC within 2	0.0%	84.2%	15.8%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=10,000</i>	Lowest AIC	0.0%	54.7%	45.3%
	AIC within 2	0.0%	72.8%	27.2%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=40,000</i>	Lowest AIC	0.0%	55.9%	44.1%
	AIC within 2	0.0%	72.9%	27.1%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

First, consider the fit indices plots for the structure A weak pattern simulations in Figure 5.4. For the smaller sample sizes ($N = 500$ and $N = 1,000$) the 95% coverage intervals for both the existing and proposed fit indices under consideration overlapped considerably across the 3-, 4-, and

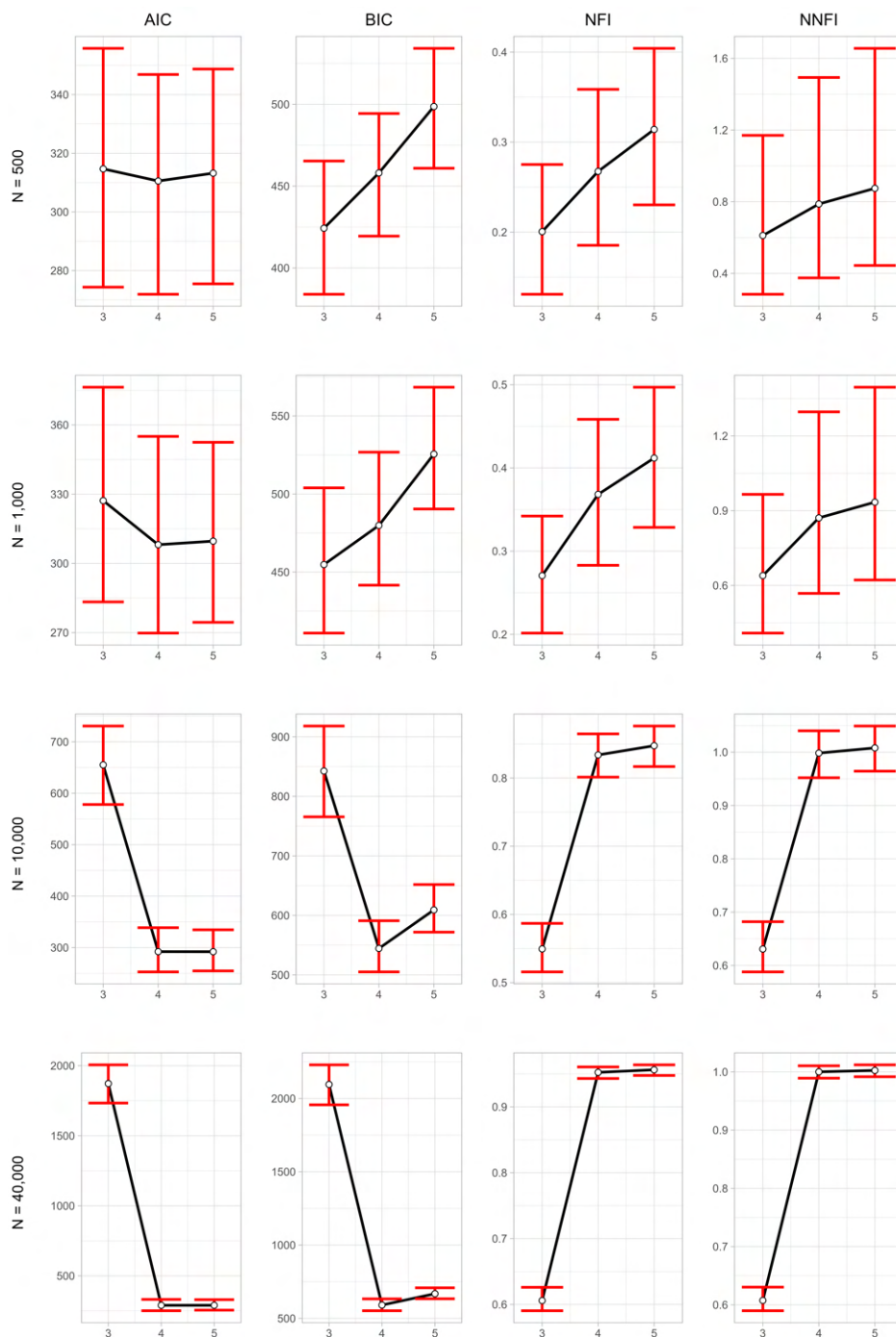


Figure 5.4: Distributions of fit indices values for the **structure A weak pattern** simulations with **equal class sizes**. *Notes:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

5-class models. While the relative change in mean value for the AIC, NFI and NNFI appeared somewhat larger between the 3- and 4-class models compared to the 4- and 5-class models, the difference was not as clear cut as for the previous set of simulations. The AIC, BIC, NFI, and NNFI thus did not visually distinguish between the correctly specified and mis-specified models as clearly for the weak pattern simulations as for the strong pattern simulations with smaller sample sizes. It is also noted that, for the smaller sample sizes, the NFI values were lower for the weak pattern simulations compared to the strong pattern simulations. For example, the maximum NFI value for the 5-class structure A weak pattern simulations with $N = 500$ was around 0.4, compared to $\text{NFI} > 0.9$ for the strong pattern equivalent. As sample size increased to $N = 10,000$ and $N = 40,000$, the “elbow” pattern that was observed in the earlier strong pattern plots reemerged for all four fit indices under consideration. The NFI values for the fitted models also increased as sample size increased.

A summary of the fit indices selection rules applied to the structure A weak pattern simulations is provided in Table 5.5. The known tendency for the BIC to select models with too few classes for smaller sample sizes was evident in these simulations, with both BIC criteria incorrectly favouring the under-specified 3-class models when $N = 500$ and $N = 1,000$. The AIC criteria also performed poorly for these smaller sample sizes, sometimes overestimating and sometimes underestimating the true number of classes. The performance of the BIC criteria improved markedly as sample size increases to $N = 10,000$ and $N = 40,000$, selecting the correct number of classes in 100% of cases. While the AIC criteria ceased to sometimes favour the 3-class solution as sample size increased, the issue of sometimes favouring the over-specified 5-class model persisted, with the “lowest AIC” criteria choosing the over-specified model in

around 46% of cases for the two larger sample sizes.

Table 5.5: Percentage of **structure A weak pattern** simulated data sets with **equal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	24.8%	56.2%	19.0%
	AIC within 2	38.3%	52.8%	8.9%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	6.6%	10.3%	83.1%
<i>N=1,000</i>	Lowest AIC	0.8%	70.2%	29.0%
	AIC within 2	2.1%	82.0%	15.9%
	Lowest BIC	99.2%	0.8%	0.0%
	BIC within 2	99.6%	0.4%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	3.0%	24.6%	72.4%
<i>N=10,000</i>	Lowest AIC	0.0%	53.9%	46.1%
	AIC within 2	0.0%	70.6%	29.4%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	0.0%	98.4%	1.6%
<i>N=40,000</i>	Lowest AIC	0.0%	54.3%	45.7%
	AIC within 2	0.0%	72.1%	27.9%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

Note that, if none of the fitted models reached the NFI cut-off of 0.90 or the NNFI cut-off of 0.95, the largest fitted model (with 5 classes) was selected as the optimal model. Thus, the 5 class column in Table 5.5 indicates the percentage of simulations for which the selection rules suggested *at least* 5

classes. The $NFI > 0.90$ criteria suggested selecting more than 4 classes for 100% of the simulated datasets for $N = 500$, $N = 1,000$, and $N = 10,000$, though it identified the correct number of classes in 100% of cases for the largest sample size ($N = 40,000$). The $NNFI > 0.95$ criteria also tended to overestimate the number of classes for the smaller sample sizes, and did not select the 4-class solution with 100% accuracy until sample size was increased to $N = 40,000$ (though 98.4% accuracy was achieved when $N = 10,000$). These findings suggest that the “rule-of-thumb” cut-off values established to guide NFI and NNFI interpretation in SEM are not appropriate for some LCA models.

Briefly, very similar findings were observed when structure B was paired with the weak item response pattern to generate the simulation data rather than structure A. The results for this set of simulations are summarised in Figure 5.5 and Table 5.6, and closely mirror the results for the structure A weak pattern simulations summarised in Figure 5.4 and Table 5.5.

Overall, both the commonly used AIC and BIC measures and the proposed NFI and NNFI measures for LCA performed worse when data was simulated using the weak item response pattern (meaning the underlying model has poorer class homogeneity) compared to the strong pattern paired with the same model structure. While the performance of the NFI, NNFI, and BIC for these simulations noticeably improved as sample size increased, the AIC continued to have issues.

Broadly, the width of the 95% coverage intervals for the NFI and NNFI visibly narrowed as sample size increased across all data generation models. It was also observed that the NFI and NNFI means increased as sample size was increased, particularly for the weak pattern simulations. From this observation, one might expect that NFI and NNFI values calculated for

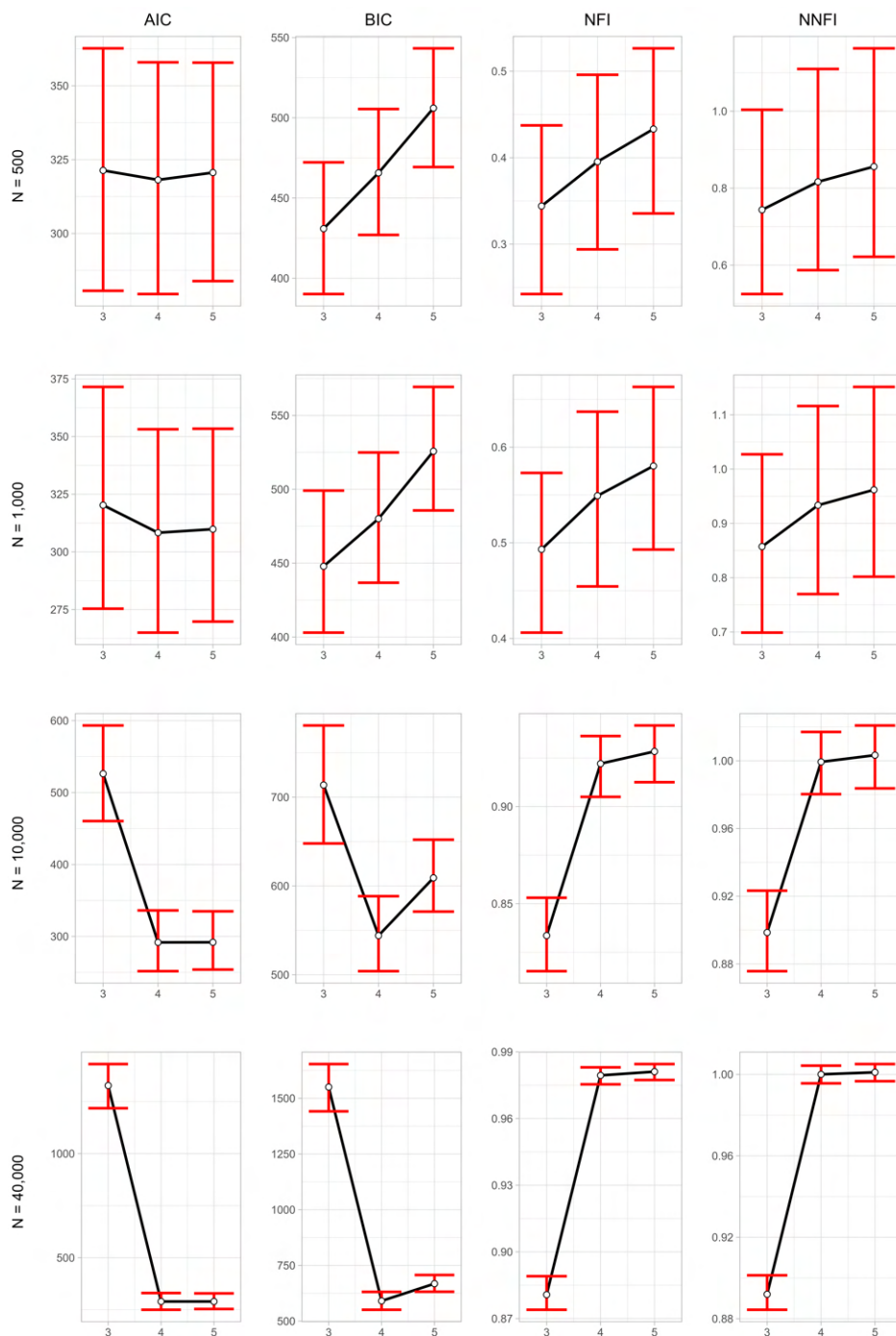


Figure 5.5: Distributions of fit indices values for the **structure B weak pattern** simulations with **equal class sizes**. *Notes:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

Table 5.6: Percentage of **structure B weak pattern** simulated data sets with **equal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	27.6%	52.9%	19.5%
	AIC within 2	42.2%	49.9%	7.9%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	5.2%	7.8%	87.0%
<i>N=1,000</i>	Lowest AIC	4.8%	65.4%	29.8%
	AIC within 2	8.7%	75.6%	15.7%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	11.9%	27.0%	61.1%
<i>N=10,000</i>	Lowest AIC	0.0%	54.1%	45.9%
	AIC within 2	0.0%	73.3%	26.7%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	99.4%	0.6%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=40,000</i>	Lowest AIC	0.0%	55.3%	44.7%
	AIC within 2	0.0%	72.7%	27.3%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

models fitted to larger samples will generally be larger in practice. Notably, the mean NFI values for the correct four-class LCA models approached the maximum value of one as sample size increased for all structure and class size combinations considered.

The implications of these results are explored further in the following Discussion subsection.

5.5.1.3 Discussion

Simulation study 1 provided numerous useful insights into how the proposed NFI and NNFI measures for LCA might behave and be used in practice, and how the proposed indices may compare to some commonly used LCA model selection tools. This preliminary study also served to highlight serious limitations of the proposed RMSEA analogue for LCA. Due to these limitations, it was concluded that the proposed RMSEA adaptation would not be beneficial as a class enumeration tool for LCA. The remainder of this discussion thus focuses on the NFI and NNFI.

It appears that whether structure A or B was used and whether class sizes were equal or unequal for the data generation models did not strongly influence the behaviour and performance of the fit indices under consideration. Item response pattern (strong/weak) appeared to be a much stronger driver of fit index performance. This trend is likely due to the data generation models for the strong pattern simulations having higher class separation and class homogeneity than for the weak pattern simulations. As both class separation and class homogeneity are desirable traits that reduce noise and simplify the recovery of the underlying structure by the fitted model, the task of selecting an appropriate number of classes was likely more straightforward for the existing and proposed fit indices for the strong pattern simulations compared to the weak pattern simulations.

Both the summary plots and selection rules for the NFI and NNFI applied to the strong pattern simulations clearly indicated the correct 4-class solution as the optimal model in all cases. While the BIC performed just as well,

the NFI and NNFI outperformed the AIC in some cases where the AIC chose the over-specified 5-class model instead. However, for the weak pattern simulations, all four fit indices had issues for the smaller sample sizes ($N = 500$ and $N = 1,000$). While the BIC and AIC criteria were susceptible to under- and over-estimating the required number of classes, the cut-off values commonly applied to interpret the NFI and NNFI in SEM proved too high to identify the correct number of classes for these simulations. This finding suggests that the rule-of-thumb cut-offs used for these indices in SEM are not directly transferable to LCA.

Furthermore, the value of the NFI and NNFI means for the models with different numbers of classes varied considerably depending on both the population model underlying the simulations and the sample size. For example, the mean NFI for the under-specified 3-class models fitted to the structure A strong pattern simulations with equal class sizes and $N = 1,000$ was above 0.60 (Figure 5.2), while the mean NFI for the correctly-specified 4-class models fitted to the structure A weak pattern simulations with equal class sizes and $N = 1,000$ was below 0.40 (Figure 5.4). Thus, it may not be appropriate to suggest using rule-of-thumb cut-off values at all to indicate adequate model fit for these indices, as is common practice for their SEM analogues.

While the NFI and NNFI values calculated for each model varied considerably depending on the underlying LCA structure and the sample size, most NFI and NNFI plots displayed an “elbow” shape whereby there was a larger increase in mean NFI/NNFI values between the under-specified three-class LCA models and correctly specified four-class LCA models, and a smaller increase in mean NFI/NNFI values between the correct four-class and over-specified five-class LCA models. This elbow was particularly apparent for all strong pattern data generation models in Figures 5.2 and 5.3, and for the

weak pattern data generation models when $N \geq 10,000$ in Figures 5.4 and 5.5. This behaviour may indicate potential to introduce an “elbow” rule for interpreting the NFI and NNFI in LCA rather than using rule-of-thumb cut-offs. The application of such an elbow rule is explored in simulation study 2 in Subsection 5.5.2.

Overall, the NNFI behaved very similarly to the NFI across the various structure, pattern, class size and sample size combinations. One difference was that there was more variability (wider 95% coverage intervals) in NNFI values for the LCA models fitted to the weak pattern simulations with smaller sample sizes in Figures 5.4 and 5.5. This may suggest that the NNFI is less reliable for smaller sample sizes. For simulation study 2 attention is thus focused on the NFI rather than the NNFI.

This preliminary simulation study provided some evidence that the proposed NFI and NNFI measures may perform as well as the AIC and BIC, and sometimes even better than the AIC, when used for LCA class enumeration in some circumstances. The study also helped to rule out the potential of an RMSEA measure and the use of the SEM rule-of-thumb cut-offs for the NFI and NNFI in LCA. However, the study is limited by considering only a small number of possible data generation structures and serves only to scratch the surface of investigating the potential utility of the proposed indices. Building on the findings of this first simulation study, a second simulation study was conducted to further explore the behaviour of the indices under a different LCA model structure. This structure was designed to reflect the situation that is of particular interest, where a smaller number of classes than the true number of classes provides a satisfactory approximation of the population structure. Details of this second simulation study, and how its design was informed by the first simulation study, are included in Subsection 5.5.2.

5.5.2 Simulation study 2

As described in the previous subsection, simulation study 1 used a more standard study design whereby there was one objectively correct solution arising from four unique and distinctive latent classes in the data generation models. Setting the simulation study up in this way provided a valuable preliminary overview of how the proposed fit indices might behave in practice when used for LCA rather than SEM, and allowed comparisons to be drawn to the common behaviours of the AIC and BIC when applied in LCA to identify the “correct” number of classes.

However, the motivation for proposing the adaption of these SEM fit indices for LCA was not to identify the “correct” number of classes in relatively straightforward models such as those used in simulation study 1. The goal for these fit indices is to address a common class enumeration issue that arises when fitting LCA models, particularly with larger sample sizes. As detailed in Subsection 5.2.2, the traditional methods for determining an appropriate number of classes (such as examining the AIC and BIC) will often continue to suggest a larger number of classes, even when such a model may not be interpretable or identifiable. Introducing these fit indices to LCA aims to help identify an interpretable LCA solution with a smaller number of classes that provides a reasonable fit to the data.

Thus, a more nuanced study design was adopted for simulation study 2 to reflect this specific challenge of interest. Specifically, a 5 class data generation model was designed that could be adequately described by a smaller 4 class model if two practically similar classes were combined. Then, how well each fit index performed at selecting the simpler model was assessed. The design for simulation study 2 was also adapted based on the findings from simulation study 1. Further details on the design of the second simulation study, and

how it was influenced by the results of the first simulation study, are provided under the below Design heading.

5.5.2.1 Design

A data-generating model with 5 latent classes and 6 dichotomous manifest indicators was used for simulation study 2. The latent class structure is presented in Table 5.7.

Table 5.7: Data-generating model parameters simulation study 2.

	Class				
	1 (Low)	2 (Medium)	3 (High)	4 (Atypical)	5 (Slightly Different)
<i>Latent Class Membership Probability</i>	.30	.25	.15	.15	.15
<i>Indicator</i>	<i>Item-Response Probability for 'Yes' Response</i>				
1	.8	.9	.8	.8	.7
2	.8	.9	.8	.2	.7
3	.2	.9	.8	.8	.7
4	.2	.9	.8	.2	.7
5	.2	.1	.8	.8	.3
6	.2	.1	.8	.2	.3

Notes: Class 5 is conceptualized as practically similar to Class 2.

As the normed and non-normed fit indices are motivated by scenarios where the population has a wide range of classes but a parsimonious model fits “well enough,” the simulation study was designed to reflect such a scenario. Specifically, the data-generating model had 5 latent classes, but Class 5 was small and theoretically similar to the larger Class 2. Thus, a 4-class model could arguably describe the data “well enough,” for example by merging Classes 2 and 5 together.

It was of interest to investigate the performance of the proposed fit indices in a scenario where (a) for small sample sizes, the traditional class

enumeration approaches (AIC, BIC, and BLRT) suggest 4 classes, but (b) for large sample sizes, the traditional approaches suggest 5 (or more) classes, detecting the statistically (but not practically) significant differences in item-response probabilities between Classes 2 and 5. Thus, similar to simulation study 1, a wide range of sample sizes were considered ($N = 500; 5,000; 10,000; 30,000; 100,000$) to investigate how well the proposed indices perform for different sample sizes. The maximum sample size was increased up to 100,000 for this more refined simulation study to reflect how large some cohort studies have become in practice. Real-world data examples include the Survey of Health, Ageing and Retirement in Europe (SHARE), an older adult cohort study with nearly 140,000 participants across Europe and Israel as of Wave 7 in 2019 (Bergmann et al., 2019); and the UK Biobank, a cohort study of over half a million UK citizens aged 40-69 at baseline (Sudlow et al., 2015).

As in simulation study 1, 1000 replication samples were generated based on the model parameters in Table 5.7 for each sample size. Note that the number of manifest items was reduced from 8 in simulation study 1 down to 6 in simulation study 2. As removing 2 items did not compromise the goal of the study design, this choice was made to reduce computational time when fitting LCA models to the simulated data. Further, as varying class membership probabilities from equal to unequal in simulation study 1 did not appear to affect the behaviour of the proposed indices, class sizes were not varied in simulation study 2. Class membership probabilities were held constant at 0.30, 0.25, 0.15, 0.15, and 0.15 for Classes 1, 2, 3, 4, and 5 respectively.

While simulation study 1 established that a rule-of-thumb cut-off may not be an appropriate approach for interpreting the proposed NFI and NNFI fit indices with LCA, potential for implementing an “elbow rule” for interpreting these fit indices instead was noted. Thus, models with a wider range of class

numbers (1-6 classes) were fitted for simulation study 2 to better facilitate the identification of such an elbow. Again, to ensure identification of the maximum likelihood solution, 100 sets of random starting values were used for each model in each replication sample.

The NFI, NNFI, AIC, and BIC were computed for all fitted models in simulation study 2. The BLRT was also computed for all models. Note that the BLRT was not calculated for the preliminary investigation in simulation study 1 as it is very computationally intensive to run. Variations of the AIC and BIC information criteria such as the ABIC were not considered, as the ABIC is usually stricter than the AIC but less strict than the BIC (Dziak et al., 2020). Thus, the ABIC would not provide further information to address the issue of interest in this work, namely when both the AIC and BIC select a LC solution with more classes than necessary to adequately describe the structure in the data. Similar to the first simulation study, the average NFI, NNFI, AIC, and BIC were calculated and plotted along with 95% coverage intervals to visually examine each model selection tool across the true and mis-specified latent class models. The summaries of class selection for each fit index based on a number of decision rules were also reproduced, with the addition of a selection rule for the BLRT and a new elbow selection approach for the NFI (and NNFI).

The BLRT was used to select a number of classes by sequentially comparing the fit of the model with k classes to the model with $k + 1$ classes for $k = 1, 2, 3, 4, 5$. The model with the lowest value of k for which the difference in fit was not significant ($p > 0.05$) was selected.

The informal “elbow” rule was introduced for the NFI (and NNFI) based on the potential observed for such a selection rule in simulation study 1. In practice, this approach would involve the researcher examining a plot of the

fit index values calculated for the range of models with two classes up to the maximum number of classes, and selecting the number of classes from which point the relative increase in the fit index value associated with the addition of an extra class becomes negligible. In order to automate the identification of an “elbow” for the simulation study, the following NFI rule was adopted: the model with k classes was selected over the model with $k - 1$ classes if and only if $\text{NFI}(k) - \text{NFI}(k - 1) \geq (1/2)(\text{NFI}(k - 1) - \text{NFI}(k - 2))$, thus stopping when an improvement in fit is less than half the previous improvement.

R (R Core Team, 2022) was again used for data generation and fit index calculations and LatentGOLD (Vermunt and Magidson, 2021) version 6.0 was used to fit the LCA models. Simulation study 2 is included in the technical report that forms the basis of this chapter, and so the code is available in the GitHub repository for the technical report: <https://github.com/Eva-Ryan/LCA-fit-indices>.

5.5.2.2 Results

For the second simulation study, the NNFI values were again observed to behave very similarly to the NFI values across the different simulations. For brevity, summary tables and figures for the NNFI distributions have thus been included in an appendix at the end of this thesis rather than in the main text for simulation study 2 (Appendix A).

Figure 5.6 shows the distributions of the AIC, BIC, and NFI across replication samples for models with 1-6 classes and all sample sizes considered. Visually, all of the model selection tools generally suggested 4 classes over any of the inadequate smaller models (except for the BIC with $N = 500$, which suggested 3 classes). As sample size increased, however, all of the model selection tools on average suggested that 4-6 classes provided approximately

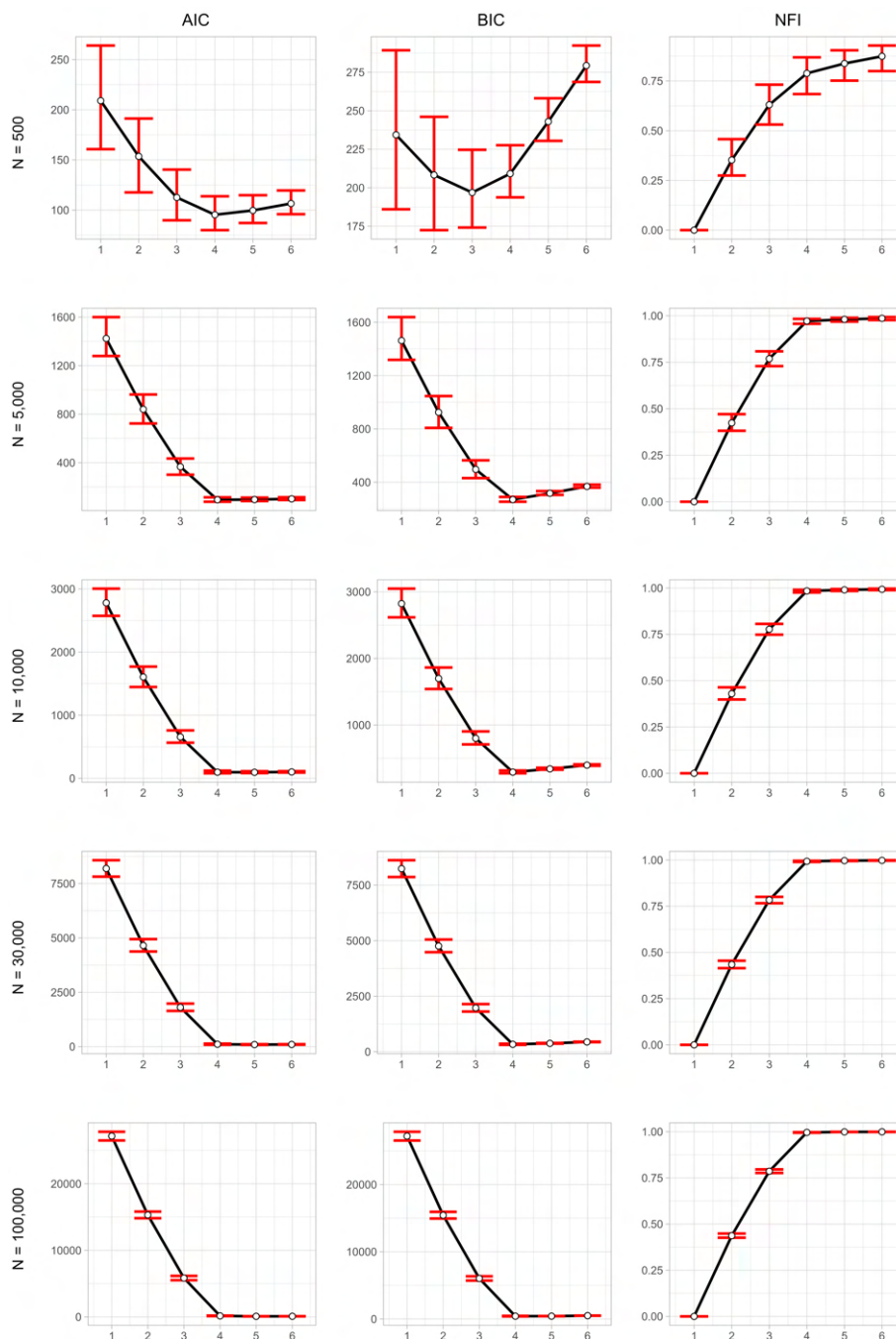


Figure 5.6: Distributions of AIC, BIC, and NFI values for models with 1-6 classes for different sample sizes. *Notes:* The x-axis includes 1-6 classes; the y-axis is the fitted value of the given model selection tool. The true number of classes specified by the simulation design is 5, but 4 classes are considered adequate for a practically reasonable answer. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

equal balance of fit and parsimony (AIC: $N \geq 10,000$; BIC: $N \geq 30,000$; NFI: $N \geq 5,000$).

Table 5.8 shows the percentages of replication samples in which the decision rules for the AIC, BIC, NFI, and BLRT selected each candidate number of classes across sample sizes. For $N = 500$, the BIC decision rules often suggested too few classes (3 or less); likely due to insufficient statistical power to fully distinguish the classes. To a much lesser extent, the AIC and BLRT also sometimes suggested too few classes when $N = 500$. For the NFI, applying the SEM cut-off rule resulted in too many classes (6) being selected for almost all simulations (96.6%). The elbow rule for the NFI was more likely to select an “adequate” number of classes (4) when $N = 500$, although it also selects too few classes for over a third of the simulations (34.5%). As sample size was increased to 5,000 and then to 10,000, the BIC decision rules and the NFI > 0.9 rule selected 4 classes for 100% of simulations. The NFI elbow rule also tended towards selecting 4 classes for most simulations (84.1% and 92.6% for $N = 5,000$ and $N = 10,000$ respectively). Meanwhile, the AIC and BLRT rules were split between selecting the “practically best” 4 class solution and the “correct” 5 class solution as sample size was increased to 10,000.

For $N = 30,000$, the AIC and BLRT rules mostly suggested 5 classes, whereas the BIC and NFI rules tended to suggest 4 classes. For $N = 100,000$, the AIC and BLRT suggested 5 classes and even 6 classes in some cases. While the BIC rules again selected 4 classes for a majority of simulations when $N = 100,000$, they also favoured the 5-class solution in many cases (41.9% and 37.1% of simulations respectively for lowest BIC and BIC within 2). In comparison, the NFI continued to suggest the practical 4-class solution 100% of the time for both decision rules (NFI $> .90$, NFI “elbow”) for the largest sample size.

Table 5.8: Percentage of simulated data sets in which latent class model sizes are selected by each model selection tool.

Sample Size	Tool	Number of Classes					
		1	2	3	4	5	6
<i>N=500</i>	Lowest AIC	0%	0%	1.8%	88.6%	9.4%	0.2%
	AIC within 2	0%	0%	3.7%	91.6%	4.5%	0.2%
	Lowest BIC	0.9%	14.8%	74.1%	10.2%	0%	0%
	BIC within 2	1.3%	19.4%	72.7%	6.6%	0%	0%
	NFI “elbow”	0%	4.5%	34.5%	57.8%	1.8%	1.4%
	NFI >.90	0%	0%	0%	0%	3.4%	96.6%
	BLRT	0%	0%	4.8%	90.1%	5.0%	0.1%
<i>N=5,000</i>	Lowest AIC	0%	0%	0%	67.4%	30.5%	2.1%
	AIC within 2	0%	0%	0%	79.8%	19.5%	0.7%
	Lowest BIC	0%	0%	0%	100%	0%	0%
	BIC within 2	0%	0%	0%	100%	0%	0%
	NFI “elbow”	0%	0%	15.9%	84.1%	0%	0%
	NFI >.90	0%	0%	0%	100%	0%	0%
	BLRT	0%	0%	0%	84.7%	14.9%	0.4%
<i>N=10,000</i>	Lowest AIC	0%	0%	0%	43.5%	50.6%	5.9%
	AIC within 2	0%	0%	0%	55.6%	42.0%	2.4%
	Lowest BIC	0%	0%	0%	100%	0%	0%
	BIC within 2	0%	0%	0%	100%	0%	0%
	NFI “elbow”	0%	0%	7.4%	92.6%	0%	0%
	NFI >.90	0%	0%	0%	100%	0%	0%
	BLRT	0%	0%	0%	63.0%	35.1%	1.9%

Sample Size	Tool	Number of Classes					
		1	2	3	4	5	6
<i>N=30,000</i>	Lowest AIC	0%	0%	0%	4.1%	86.7%	9.2%
	AIC within 2	0%	0%	0%	7.9%	87.3%	4.8%
	Lowest BIC	0%	0%	0%	99.9%	0.1%	0%
	BIC within 2	0%	0%	0%	100%	0%	0%
	NFI “elbow”	0%	0%	0.4%	99.6%	0%	0%
	NFI >.90	0%	0%	0%	100%	0%	0%
	BLRT	0%	0%	0%	10.8%	85.0%	4.2%
<i>N=100,000</i>	Lowest AIC	0%	0%	0%	0%	88.3%	11.7%
	AIC within 2	0%	0%	0%	0%	94.0%	6.0%
	Lowest BIC	0%	0%	0%	58.1%	41.9%	0%
	BIC within 2	0%	0%	0%	62.9%	37.1%	0%
	NFI “elbow”	0%	0%	0%	100%	0%	0%
	NFI >.90	0%	0%	0%	100%	0%	0%
	BLRT	0%	0%	0%	0%	94.2%	5.8%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. In a small number of simulations, the p-values for the BLRT were still less than 0.05 for the 5-class versus 6-class model comparisons. The percentages in the 6-class columns for BLRT thus represent the percentage of simulations for which the BLRT would suggest 6 classes or more.

5.5.2.3 Discussion

Recall that the 5-class model was the true data-generating model but the 4-class model was considered practically adequate. In general, the BIC and NFI (and NNFI) performed similarly in terms of selecting the favoured 4-class

model. However, the NFI (and NNFI) more consistently suggested this more parsimonious model when the sample size was very large, and so may have an advantage over the BIC for large N . Unsurprisingly, the AIC tended to suggest slightly larger models than either the BIC or NFI/NNFI. The BLRT also tended towards suggesting larger models as the sample size was increased.

The NFI/NNFI thus show some promise as useful class enumeration tools to aid the specific challenge this work aims to address: selecting a smaller, more interpretable model that provide a practically useful fit to the data when sample size is large, and the traditional LCA model selection tools suggest a model with an impractically large number of classes. However, while this simulation study provides some proof of concept, only one data generating model was considered and so the results cannot be generalised. More simulation studies using different model specifications are required to verify if the NFI/NNFI performs similarly for other cases. Additionally, the class membership probabilities were held constant in this simulation study as varying class membership probabilities did not appear to influence the behaviour of the proposed fit indices in simulation study 1. Given that these simulation studies considered just a few data generation models, future work should also consider how well the fit indices perform when class sizes are unequal, especially when there are some very large or very small classes. Applications to real world datasets would also be beneficial to provide further insight into how the NFI/NNFI perform in practice.

5.6 Conclusions

Scientists analysing data using LCA frequently face disagreement across traditional model selection tools when determining the optimal number of

latent classes. Additionally, these tools frequently suggest selecting a model that is too large to be identified or usefully interpreted. This issue is typically exacerbated by large sample sizes as small, practically non-significant discrepancies between the fitted model and observed data become statistically significant (Wasserstein and Lazar, 2016). Note that, although this work focuses on LCA models with categorical indicators, these issues are also common to other mixture models (Pohle et al., 2017). Direct LCA analogues of the NFI, NNFI, and RMSEA fit indices used in SEM were presented in this work as tools to aid class enumeration, particularly when sample sizes are large. While simulation studies revealed that the RMSEA is not appropriate for use in LCA, overall the NFI/NNFI were found to perform similarly to the BIC, and outperformed the BIC in choosing the most practically significant model for some very large samples simulations ($N = 100,000$). Thus, the NFI/NNFI show some potential to address the class enumeration challenge that motivated this work. However, this study is intended as a preliminary analysis to serve as a starting point for additional work on this topic, and considerable further research is required before these tools can be widely recommended with guidelines for interpretation.

The most immediate issue is that reasonable cutoffs for determining acceptability of an LCA model cannot yet be recommended for the NFI or NNFI. When used in their traditional setting of SEM with normal data, the typical approach for interpreting these fit indices is to select the smallest model for which the fit index value passes a certain threshold e.g., $NFI > 0.9$ (Bentler and Bonett, 1980). These benchmarks were established by using a range of Monte Carlo simulations to argue that, under certain assumptions, selecting a model which meets the suggested criterion will result in the true population model being selected with high probability (Hu and Bentler, 1999).

Although these cutoffs are somewhat arbitrary, they provide a common point of comparison across datasets and across scientists to eliminate obviously inadequate models. The results of the simulation studies in this work suggest that the cutoffs used in SEM are not directly transferable to LCA. Further simulation studies are required to determine if general cutoffs are feasible, or if perhaps different cutoffs dependent on factors such as sample size or number of indicator variables could be established. Applying the fit indices to LCA models fitted to real-world datasets will also provide further insight into their behaviour. Additionally, it is noted that the NFI and NNFI were observed to perform very similarly overall. Thus, in future work it may be more beneficial to focus efforts on exploring the potential of the NFI, given that the NNFI does not scale between 0 and 1 and so may be more difficult to interpret in practice.

While the simulation studies have demonstrated that the new indices have some potential as useful heuristics, as noted earlier the studies are intended as a preliminary investigation and they are subject to some limitations in practice. Replication samples were generated from reasonably simple LCA data-generation models with dichotomous indicator variables and no covariates, thus only a few possible scenarios were considered. It is difficult to recreate the situation of interest using a simulation study; that is, when candidate models are adequate, but not exactly correct. This approach is motivated by the idea that one can think of latent classes as representing zones on a possibly multidimensional continuum, rather than as a set of completely distinct groupings (for example, see Sher et al. (2011)). Monte Carlo simulations typically require an unambiguous correct answer to check against, which is in tension with the motivation to use normed and non-normed fit indices that indicate relative adequacy rather than literal truth.

While simulation study 2 was carefully designed to represent such a situation, there is no quantitative index for automatically and objectively verifying how theoretically meaningful or practically useful class distinctions are, as this depends on research questions and substantive context in practice.

There are also some technical issues to consider when conceptualising normed and non-normed fit indices for LCA. The chi-squared based G^2 fit statistic used to calculate the indices is the sum of many ratios of small quantities, which may result in the accumulation of rounding error if there are many indicators in the LCA model (Collins et al., 1993). Another potential issue is that introducing additional model selection tools, especially without clear guidelines, could increase ambiguity and confusion for researchers when trying to justify selecting one model in favour of another (Simmons et al., 2011).

A number of suggestions have been made for scientists to pursue when no LCA model with objectively good fit and good theoretical interpretability can be identified. One option would be to use one or more continuous latent factors instead of latent classes (Walters, 2011; Wendt et al., 2019). Alternatively, continuous and categorical latent variables could be combined in a factor mixture model (Lubke and Muthén, 2005). It could be argued that settling for an LCA model that is “good enough” is a distraction from pursuing such alternative models which may more effectively captures the latent construct of interest. On the other hand, even if the identified classes represent fuzzy zones within a continuum rather than completely distinct subgroups, they can still have descriptive usefulness (Sher et al., 2011). Additionally, these alternative methods rely on the availability of appropriate continuous variables that adequately measure the latent construct of interest, which might not be realistic.

The simulation studies detailed in this chapter were not intended to provide concrete recommendations or guidelines for the application and interpretation of normed and non-normed fit indices in LCA. Rather, the goal was to illustrate the ideas involved and provide proof of concept for future work. Scientists who are familiar with SEM are often curious about whether normed and non-normed fit indices are also available for use in LCA. The formal proposal of LCA analogues and preliminary simulation studies in this work are a first step in answering this question. This work may also serve to advance scientific conversations about model selection in LCA and what a latent class solution is or is not intended to represent, particularly where it is intended as an interpretable data reduction technique rather than a precise description of a population distribution. The behaviour of the proposed NFI and NNFI fit indices are further explored in the next chapter, where they are used to aid the identification of pain experience subgroups in older adults using LCA applied to HRS data.

6 Pain Experience in American Older Adults: A Latent Class Analysis

6.1 Introduction

As discussed in Chapter 5, defining a comprehensive measure of chronic pain can be a difficult task. Single measures of different aspects of pain experience, such as pain intensity or pain-related disability, are commonly used. However, it could be beneficial to identify or construct a more holistic pain experience variable that encapsulates these various dimensions of pain. LCA was proposed as a potential tool to facilitate the identification of such a pain experience variable in Chapter 5. In this Chapter, LCA will be applied to identify latent classes of pain experience among a cross-sectional sample of Health and Retirement Study (HRS) 2016 Wave participants using various categorical pain-related indicator variables.

A range of pain measures available in the HRS data are considered for inclusion as indicator variables in the LCA model developed in this chapter. As mentioned previously, pain intensity (e.g., rated as mild, moderate, or severe pain) is an important and commonly used measure to represent the level of pain suffered by an individual in terms of the feeling or sensation of the pain (Fillingim et al., 2016). However, while intensity is one important dimension

of pain, Ojala et al. (2015) highlights that intensity captures the sensation of pain only and does not fully encapsulate the experience or burden of living with pain. Further, a Task Force on chronic lower back pain convened by the National Institutes of Health Pain Consortium recommended that pain should be classified in terms of impact, which encompasses not only pain intensity but also the interference of pain with daily activities and the individual's physical function (Deyo et al., 2015). Sullivan and Ballantyne (2016) suggest that while reducing pain intensity or severity is a common goal when addressing chronic pain, pain-related disability is a more appropriate measure for pain assessment and therefore a more pertinent target to reduce pain interference. Thus, in addition to pain intensity, a holistic measure of pain experience should also account for the burden of pain in terms of disability or impact on the individual's daily activities.

Pain management strategies are another important facet of pain experience, with medication often playing a central role in pain treatment plans. One study of European adults found that almost half of chronic pain sufferers were using non-prescription, "over-the-counter" (OTC) analgesics to manage their pain, while weak opioids and strong opioids were prescribed to 23% and 5% of the sample respectively (Breivik et al., 2006). Opioid prescriptions are even more common in the United States, with an analysis of data from the National Health Interview Survey 2019-2020 estimating that 29.3% (95% confidence interval: 28.2%, 30.4%) of American adults with chronic pain had used prescription opioids in the preceding 12-month period (Zajacova et al., 2023). Both OTC and opioid medication use are included as indicator variables in the LCA model of pain experience in this study.

Pain location is also an important consideration when assessing different domains of pain, as the distribution of pain sites can heavily influence pain

impact as well as diagnosis and treatment plans (Fillingim et al., 2016). For example, the Global Burden of Disease Study conducted in 2013 found that, out of a wide range of diseases and injuries, chronic low back pain was the single greatest cause of years lived with disability across the 188 countries included in the study (Rice et al., 2016). Chronic back pain is common in older adults, with a recent systematic review reporting a pooled point estimate of 20.6% (95% confidence interval: 19.4%–21.9%) for chronic low back pain prevalence in studies of community-dwelling older adults (Wong et al., 2022). As information on this prevalent and often debilitating pain sub-type is available in the HRS dataset, self-reported back pain is also included as an indicator in the LCA model in this study.

The goal of this work is to identify distinct latent classes of pain experience among a group of American older adults who reported being troubled by pain, by developing an LCA model using a diverse range of pain indicator variables. This goal is motivated by a need for more holistic measures of pain experience that account not only for the reported level of pain intensity, but also how pain impacts the individual’s life, the type of medication used to manage pain, and pain location. It is also of interest to compare the sociodemographic backgrounds of the identified pain experience latent classes to further understanding of sociodemographic differences in pain, as also explored in Chapter 3. Model fit will be assessed using the fit indices proposed in Chapter 5.

6.2 Methods

Data

Similar to the work in Chapter 4, this study analyses data from the Health and Retirement Study (HRS; Health and Retirement Study, 2023). An overview of the HRS was provided previously in Chapter 1 and Chapter 4. This work presents a cross-sectional analysis of data from the 2016 wave of the HRS.

The goal of this study is to identify unique latent classes among older American adults who experience pain; thus, the analytic sample was restricted to only those 2016 wave participants who responded “yes” when asked “Are you often troubled with pain?”. As described previously in Chapter 3 and Chapter 4, while this question does not specify a duration it is believed to capture persistent rather than transient pain (Banks et al., 2009; Grol-Prokopczyk, 2017). The 2016 wave of the HRS was chosen for this analysis as the HRS did not collect data on participants’ pain medication use prior to this wave. Of the 8,085 participants aged 51+ who reported being “often troubled by pain” in the 2016 HRS wave, 373 (4.6%) were missing values on at least one of the pain indicator variables used to identify the LCA model in this study. Those who were missing pain data were more likely to be older, female, widowed, have a smaller household, not have a degree, be retired or not in the labour force, not have private health insurance, have a lower BMI, and have arthritis or diabetes than the analytic sample (χ^2 test or Kruskal-Wallis test p-values <0.05), but were not significantly different across the other background characteristics described later in the Covariates section. A summary of the differences in background characteristics between those with and without missing pain data is included in Table B.1 of Appendix B. As detailed later,

missing data can reduce the reliability of bivariate residuals calculated to assess the LCA local independence assumption. As the proportion of individuals with missing pain data was $< 5\%$ the bias resulting from a complete cases analysis is likely to be small. Thus, it was decided to remove participants with missing pain data to give a final analytic sample of 7,712 participants.

LCA indicator items

As described in subsection 5.2.1 of Chapter 5, an LCA model is identified using data collected on observable indicator variables believed to be caused by the unobserved latent variable of interest. A number of variables related to pain experience and pain management are used as indicator variables in this study. Pain severity among those who reported being often troubled by pain in the HRS was captured by the question “How bad is the pain most of the time: mild, moderate or severe?”. Experience of pain-related disability was ascertained by asking “Does the pain make it difficult for you to do your usual activities such as household chores or work?” with “yes” or “no” as response options. HRS participants were also asked specifically about back pain: “Have you had any of the following persistent or troublesome problems: Back pain or problems?” (yes, no). Information on over-the-counter (OTC) medication usage and opioid medication usage in the preceding 3-month period was collected using the following “yes” or “no” questions: “Over-the-counter pain medications include such things as Advil, Aleve, Tylenol, aspirin or similar medications. In the past three months have you taken any over-the-counter pain medications?” and “Another class of pain medications, called “opioids”, includes such things as Vicodin, OxyContin, codeine, morphine, or similar medications. In the past three months, have you taken any opioid pain medications?”

Covariates

After identifying a final LCA model and assigning latent class membership to the sample participants, the distribution of participant characteristics within each pain latent class is examined. A comprehensive set of relevant sociodemographic and health status variables was previously identified when choosing a confounder adjustment set for the HRS analysis detailed in Chapter 4. The same set of variables is used for the within-class descriptive statistics presented in this study, with the exception of the variable that measured the importance of religion (not available for the 2016 HRS wave).

The following sociodemographic variables are summarised for each pain latent class: age in years in 2016; sex (male, female); race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, non-Hispanic Other); marital status (married, separated/divorced, widowed, never married, other); household size; number of children; region of residence (Northeast, Midwest, South, West); urbanicity (urban, suburban, ex-urban/rural); highest level of education (no degree, high school diploma, 4-year college degree, graduate degree); household wealth quartile; employment status (employed, unemployed, retired, not in labour force); food security (“In the last 2 years, have you always had enough money to buy the food you need?” yes, no); veteran status (yes, no); and health insurance type (uninsured, any private insurance, public insurance only). Additionally, the distributions of the following health status variables are summarised for each pain latent class: active cancer (yes, no); diabetes (yes, no); chronic lung disease (yes, no); angina (yes, no); stroke (yes, no); heart condition (yes, no); arthritis (yes, no); BMI category (underweight: $BMI < 18.5 \text{ kg/m}^2$, normal weight: $18.5 \text{ kg/m}^2 \leq BMI < 25 \text{ kg/m}^2$, overweight: $25 \text{ kg/m}^2 \leq BMI < 30 \text{ kg/m}^2$, obese 1: $30 \text{ kg/m}^2 \leq BMI < 35 \text{ kg/m}^2$, obese 2: $35 \text{ kg/m}^2 \leq BMI < 40 \text{ kg/m}^2$, obese 3: $BMI \geq 40 \text{ kg/m}^2$); smoking status (never smoker, former smoker, current smoker); and mental health

status as measured by the 8-item version of the Center for Epidemiological Studies-Depression Scale (on a scale of 0-8, with higher scores indicating more depressive symptoms) (Karim et al., 2015).

Analysis

LCA is used to identify distinct pain experience latent classes using the five pain experience indicator variables described above. The standard LCA model is characterised by two sets of parameters: latent class membership probabilities, which estimate the size of each latent class; and item response probabilities, which estimate the class-specific probabilities of selection for each category of each indicator variable. Local independence is a central assumption of LCA, meaning that the observed indicator variables used to estimate the model are assumed to be independent conditional on latent class membership (Collins and Lanza, 2010). If not appropriately accounted for, breaches of this assumption can result in biased model parameter estimates (Visser and Depaoli, 2022). A more detailed technical overview of the LCA model was presented previously in Chapter 5 (section 5.2) of this thesis.

Candidate LCA models with between 2 and 6 latent classes are fitted to aid the selection of an optimal number of latent classes in this study. As described in subsection 5.2.2 of Chapter 5, a range of information criteria and likelihood ratio-based significance tests are typically considered to identify the candidate LCA model with the most parsimonious fit. In this work, candidate models are compared using standard measures including the Akaike's information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), and the bootstrapped likelihood ratio test (BLRT; McLachlan and Peel, 2000). Note, 100 replicate samples are generated when carrying out the BLRT for each model. Typically, the candidate model with the lowest AIC/BIC is

selected as the optimal model. Meanwhile, a BLRT p-value > 0.05 suggests that the candidate model does not provide a significant improvement in model fit when compared to the less complex model with one fewer latent classes, and so a more parsimonious model should be chosen. However, as discussed in Chapter 5, class selection using these traditional methods is often not straightforward. The methods can suggest different numbers of latent classes, or suggest infeasibly complicated models over more parsimonious alternatives, particular when sample sizes are large. Thus, the use of the normed fit index (NFI) and non-normed fit index (NNFI) analogues for LCA proposed in Chapter 5 are also considered. The preliminary investigation in Chapter 5 suggested that the NFI and NNFI may be best interpreted by looking for an “elbow” in plots of the indices by number of latent classes to identify the most parsimonious candidate model.

Identifiability of the candidate model is another important consideration when selecting the optimal number of latent classes. As LCA model parameters are estimated by maximising a likelihood function, it is recommended to run the maximum likelihood algorithm with a range of different sets of random starting values to reduce the risk of selecting a local maximum solution (Collins and Lanza, 2010). If most of the sets of starting values converge to the same solution, this increases confidence that the global maximum solution has been identified for the final model. 100 sets of random starting values are used when estimating each candidate model in this work. The interpretability of each candidate model is also considered when selecting the final model.

Bivariate residuals (BVRs) are calculated for each candidate model to detect potential breaches of the local independence assumption. BVRs measure residual correlations between pairs of indicator variables using the Pearson test statistic, with high values suggesting local dependences. As the

Pearson test statistic rarely conforms to a chi-squared distribution in practice when used for BVRs, a parametric bootstrap procedure is employed to estimate p-values when testing if a BVR is significantly large (Visser and Depaoli, 2022). 500 bootstrap replicates are generated when calculating BVR p-values in this study. Note, BVRs are not class specific, rather they measure the overall residual correlations across latent classes.

After selecting a final model, each individual in the sample is assigned to the latent class for which they had the highest posterior probability of membership, calculated using Bayes theorem (Collins and Lanza, 2010). Descriptive statistics are then calculated for each latent class. Numeric sociodemographic and health-related covariates are summarised using the median and inter-quartile range, while categorical variables are summarised using counts and percentages.

All analyses are carried out using R (R Core Team, 2022). The tidyverse (Wickham et al., 2019) collection of packages are used for data cleaning. The poLCA (Linzer and Lewis, 2011) and poLCA.extras (Oberski, 2023) packages are used to fit the LCA models and calculate BVRs respectively. Plots are created using the ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017), and ggpubr (Alboukadel, 2022) packages. The tableone package (Yoshida and Bartel, 2022) is used to generate descriptive statistics for each latent class. R code for this analysis is available in the following GitHub repository: <https://github.com/Eva-Ryan/hrs-pain-lca>. A 5% significance level has been used for all tests.

6.3 Results

Descriptive statistics for the pain experience indicator variables used to identify the LCA model are presented in Table 6.1. The most commonly reported pain severity is moderate pain (52.4%), with severe pain being least common (19.4%). A majority of the sample reported pain-related disability (64.9%). 76.1% of the sample reported having taken OTC pain medications in the past 3 months, while 28.1% reported having taken opioid pain medications in the past 3 months. A large proportion of the sample (71.7%) reported back pain.

Table 6.1: Descriptive statistics for the LCA indicator variables.

Variables	Count (%)
<i>Pain severity</i>	
Mild	2,172 (28.2%)
Moderate	4,043 (52.4%)
Severe	1,497 (19.4%)
<i>Pain-related disability</i>	
Yes	5,005 (64.9%)
No	2,707 (35.1%)
<i>Taken OTC medication in last 3 months</i>	
Yes	5,870 (76.1%)
No	1,842 (23.9%)
<i>Taken opioid medication in last 3 months</i>	
Yes	2,169 (28.1%)
No	5,543 (71.9%)
<i>Back pain</i>	
Yes	5,531 (71.7%)
No	2,181 (28.3%)

Note: OTC = over-the-counter.

6.3.1 Model selection

A summary of model fit measures for the candidate models with 2-6 latent classes is presented in Table 6.2. Plots of the AIC, BIC, NFI, and NNFI values for each candidate model are included in Figures 6.1(a), 6.1(b), 6.1(c), and 6.1(d) respectively. The lowest AIC value (AIC = 83.3) is achieved by

the 5-class model. The BLRT also selects the 5-class model, as the 5-class model appears to give a significantly better fit than the 4-class model (p -value = $0.03 < 0.05$), but the addition of a 6th latent class does not significantly improve model fit (p -value = $0.15 > 0.05$). However, the lowest BIC value (BIC = 240.3) is achieved by the 3-class model. Observing the NFI and NNFI plots in Figures 6.1(c) and 6.1(d), the elbows in the plots also both suggest the 3-class model. Prior research has shown that the AIC tends to select a model with too many latent classes when sample sizes are large (Nylund et al., 2007). This AIC behaviour was also observed in the LCA simulation studies detailed in Chapter 5 of this thesis. The BLRT also tended to select too many latent classes for some of the large sample size simulations, while the BIC, NFI, and NNFI consistently selected the most parsimonious model. Additionally, the 3-class candidate model appears to be well-identified, with 99% of random sets of starting values converging to the same maximum likelihood solution. Meanwhile, only 21% of the 100 random sets of starting values give the same results for the 5-class model. For this reason, the 3-class model is selected as the final model.

Table 6.2: Summary of fit measures for candidate LCA models with 2-6 classes.

Number of classes	G ²	df	P	AIC	BIC	NFI	NNFI	% same ML solution	BLRT p-value
2	226.5	34	13	252.5	342.8	0.91	0.91	100%	<0.001
3	61.3	27	20	101.3	240.3	0.98	0.98	99%	<0.001
4	32.6	20	27	86.6	274.2	0.99	0.99	16%*	<0.001
5	15.3	13	34	83.3	319.6	0.99	1.00	21%	0.030
6	4.3	6	41	86.3	371.2	1.00	1.00	6%*	0.150

Notes: df = degrees of freedom; P = number of parameters in the fitted model; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; NFI = Normed Fit Index; NNFI = Non-Normed Fit Index; % same ML solution = percentage of the 100 sets of starting values for which the same best maximum likelihood solution was found.

*EM algorithm did not converge to a maximum likelihood solution within 5,000 iterations for some sets of starting values.

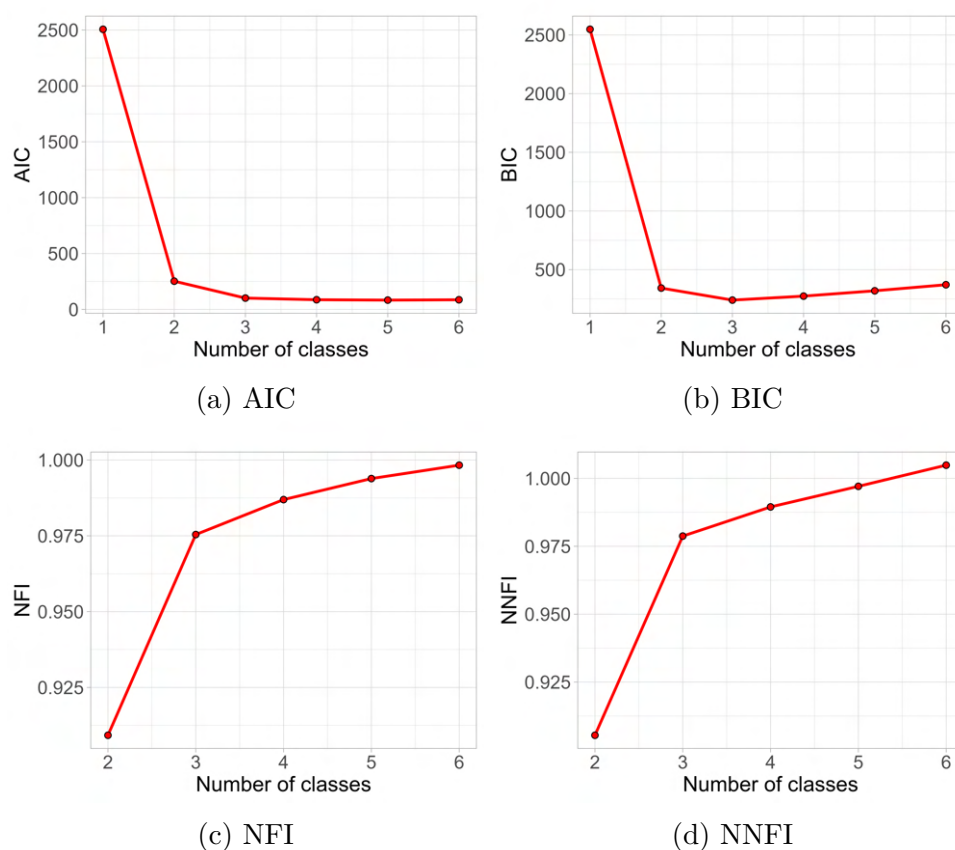


Figure 6.1: Fit indices plots for the candidate LCA models with 2-6 latent classes.

Bivariate residuals for the 3-class LCA model are presented in Table 6.3. Bootstrap p-values for each pair of variables are included in brackets. The p-values are not significant at the 0.05 level for most pairs of variables. However, the p-values for the pair back pain and OTC medication (p-value = 0.01) and the pair back pain and opioid medication (p-value = 0.04) are significant. This may suggest some residual local dependence between these pairs of variables within classes. While selecting an alternative model with more latent classes would reduce these BVRs to negligible levels, it was shown in Table 6.2 that models with more than 3 classes are not well identified based on the percentage of sets of random starting values that converge to the same optimal solution. Thus, the 3-class model will remain the final model for further analysis. Note,

Table 6.3: Bivariate residuals for the candidate 3-class LCA model.

Variables	Pain severity	Pain-related disability	OTC medication	Opioid medication
Pain-related disability	0.79 (1.00)			
OTC medication	1.03 (0.95)	0.05 (1.00)		
Opioid medication	0.57 (1.00)	0.31 (0.88)	0.03 (1.00)	
Back pain	0.69 (0.60)	0.04 (1.00)	5.53 (0.01)	4.64 (0.04)

Notes: Bootstrap p-values are included in brackets. OTC = over-the-counter.

summary tables of BVRs for the 2, 4, 5, and 6-class candidate models are included in Appendix B for reference.

6.3.2 Latent classes

The latent class membership probabilities and item response probabilities for the 3-class LCA model are summarised in Table 6.4. These parameter estimates are also visualised in Figure 6.2. Class 1 and Class 3 are estimated to be of similar size, each with latent class membership probabilities of 0.41. Class 2 is estimated to be smaller, with a membership probability of 0.18. Class 1 is characterised by high probabilities of back pain (0.80) and pain-related disability (0.92). Moderate pain is the most common pain severity in Class 1, with an item response probability of 0.63. Members of Class 1 are likely to have taken OTC pain medications in the three months prior to interview (item response probability = 0.82), but only 21% of individuals in this class are estimated to have taken opioid pain medications in the same time period. The label “**High-Impact Pain (OTC)**” will thus be used for Class 1.

Similar to Class 1, Class 2 is also characterised by high probabilities of pain-related disability (0.95) and back pain (0.91). However, OTC pain medication use is less common in Class 2 (item response probability = 0.56). Meanwhile, opioid medication use is much more common compared to Class 1, with 81%

of Class 2 members estimated to have taken opioids. Severe pain is also more common in Class 2 compared to Class 1, with 53% of Class 2 members estimated to experience severe pain. Based on the difference in pain severity and types of pain medication usage compared to Class 1, Class 2 will be labelled “**High-Impact Severe Pain (Opioids)**”.

Overall, Class 3 is characterised by much lower probabilities of pain-related disability (0.25), back pain (0.55), and severe pain (0.02) compared to the other latent classes. Mild pain is the most common pain severity in Class 3, affecting an estimated 52% of class members. While OTC medication use was common in Class 3 (item response probability = 0.79), only 12% of members are estimated to have taken opioid medication. Due to the comparatively mild severity and disabling effects of pain in Class 3, this class will be labelled “**Low-Impact Mild Pain**”.

Table 6.4: Summary of the fitted 3-class LCA model.

	Classes		
	1	2	3
Class membership probabilities	0.41	0.18	0.41
Indicator variables	Item response probabilities		
<i>Pain severity:</i>			
Mild	0.15	0.03	0.52
Moderate	0.63	0.44	0.46
Severe	0.22	0.53	0.02
<i>Pain-related disability (Yes)</i>	0.92	0.95	0.25
<i>Back pain (Yes)</i>	0.80	0.91	0.55
<i>Taken OTC pain medication (Yes)</i>	0.82	0.56	0.79
<i>Taken opioid medication (Yes)</i>	0.21	0.81	0.12

Note: OTC = over-the-counter.

Sample participants are each assigned to the latent class for which they had the highest posterior probability of membership. The average maximum posterior probability is 0.78, suggesting low uncertainty in latent class assignment. Table 6.5 summarises the distribution of sociodemographic

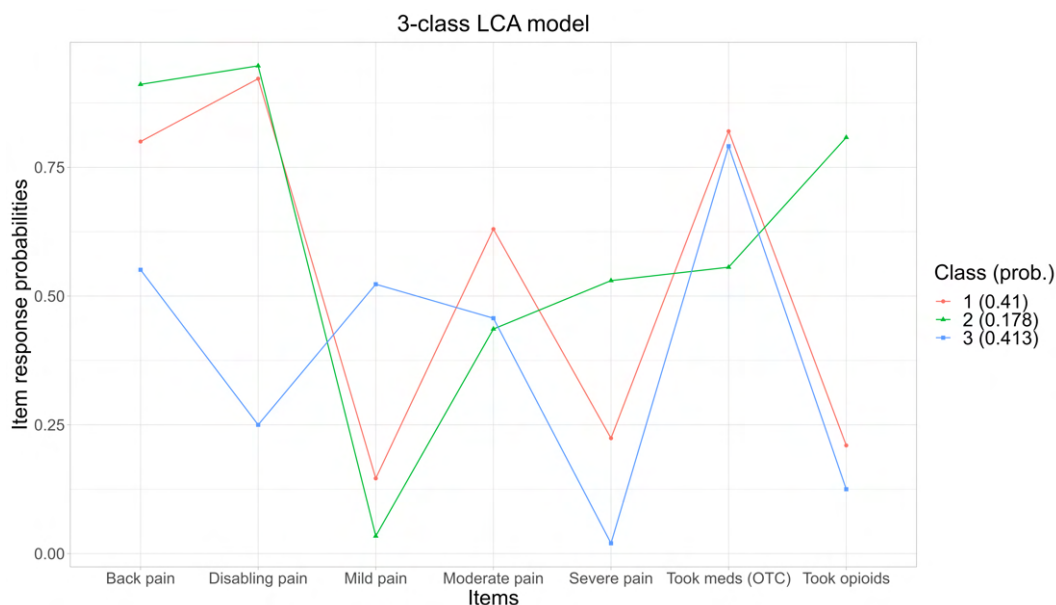


Figure 6.2: Summary of the estimated item response probabilities conditional on latent class membership for the 3-class LCA model.

and health characteristics for the participants in each latent class. The median age is lowest in the **High-Impact Severe Pain (Opioids)** class (62.0), with an inter-quartile range (IQR) of 5.0. Median age is highest in the **Low-Impact Mild Pain** class (66.0), however the IQR is also larger (8.0). The **Low-Impact Mild Pain** class has the lowest proportion of female participants (56.6%), while the **High-Impact Severe Pain (Opioids)** class has the highest proportion of females (66.8%). Median household size (2.0) and median number of children (3.0) are the same for all three latent classes. The **Low-Impact Mild Pain** class has the highest proportion of White (60.9%), married (58.7%), and employed (42.2%) members, and the highest proportion of members with 4-year (14.6%) or graduate degrees (10.9%), with any private health insurance (55.0%), with food security (91.8%), in the top wealth quartile (26.7%), and with veteran status (15.4%). On the other hand, the **High-Impact Severe Pain (Opioids)** class has the highest proportion of members

that are Black (28.1%), separated/divorced (31.1%), not in the labour force (54.9%), in the lowest wealth quartile (45.2%), and who have public health insurance only (64.0%).

Table 6.5: Descriptive statistics for each of the three classes in the final LCA model.

Variable	Class		
	Class 1 High-Impact Pain (OTC) (n = 3,787)	Class 2 High-Impact Severe Pain (Opioids) (n = 1,038)	Class 3 Low-Impact Mild Pain (n = 2,887)
<i>Age in years</i>	64.0 (8.0)	62.0 (5.0)	66.0 (8.0)
<i>Sex: Female</i>	2,442 (64.5%)	693 (66.8%)	1,633 (56.6%)
<i>Race/ethnicity</i>			
White (non-Hispanic)	2,152 (56.9%)	516 (49.7%)	1,756 (60.9%)
Black (non-Hispanic)	823 (21.7%)	292 (28.1%)	543 (18.8%)
Hispanic	610 (16.1%)	167 (16.1%)	449 (15.6%)
Other (non-Hispanic)	200 (5.3%)	63 (6.1%)	136 (4.7%)
<i>Marital status</i>			
Married	1,758 (46.5%)	421 (40.6%)	1,691 (58.7%)
Separated/divorced	941 (24.9%)	323 (31.1%)	509 (17.7%)
Widowed	715 (18.9%)	201 (19.4%)	479 (16.6%)
Never married	355 (9.4%)	91 (8.8%)	192 (6.7%)
Other	11 (0.3%)	2 (0.2%)	11 (0.4%)
<i>Household size</i>	2.0 (2.0)	2.0 (2.0)	2.0 (1.0)
<i>Number of children</i>	3.0 (2.0)	3.0 (2.0)	3.0 (2.0)

Variable	Class		
	1 High-Impact Pain (OTC) (n = 3,787)	2 High-Impact Severe Pain (Opioids) (n = 1,038)	3 Low-Impact Mild Pain (n = 2,887)
<i>Region</i>			
Northeast	546 (14.4%)	131 (12.6%)	431 (15.0%)
Mid-west	769 (20.3%)	198 (19.1%)	580 (20.1%)
South	1,698 (44.9%)	486 (46.9%)	1,247 (43.3%)
West	767 (20.3%)	222 (21.4%)	621 (21.6%)
<i>Urbanicity</i>			
Urban	1,908 (50.6%)	541 (52.2%)	1,589 (55.2%)
Suburban	819 (21.7%)	223 (21.5%)	509 (17.7%)
Ex-urban/rural	1,047 (27.7%)	272 (26.3%)	781 (27.1%)
<i>Education level</i>			
No degree	821 (23.1%)	244 (25.2%)	473 (17.1%)
High school degree	2,133 (60.1%)	604 (62.3%)	1,586 (57.4%)
4-year college degree	374 (10.5%)	85 (8.8%)	403 (14.6%)
Graduate degree	223 (6.3%)	37 (3.8%)	301 (10.9%)
<i>Wealth quartile</i>			
Q1	1,319 (34.8%)	469 (45.2%)	662 (22.9%)
Q2	1,064 (28.1%)	271 (26.1%)	709 (24.6%)
Q3	771 (20.4%)	186 (17.9%)	745 (25.8%)
Q4 (wealthiest)	632 (16.7%)	112 (10.8%)	771 (26.7%)
<i>Employment status</i>			
Employed	983 (26.2%)	122 (11.9%)	1,210 (42.2%)

Variable	Class		
	1 High-Impact Pain (OTC) (n = 3,787)	2 High-Impact Severe Pain (Opioids) (n = 1,038)	3 Low-Impact Mild Pain (n = 2,887)
Unemployed	170 (4.5%)	26 (2.5%)	106 (3.7%)
Retired	1,380 (36.8%)	316 (30.7%)	1,200 (41.8%)
Not in labour force	1,221 (32.5%)	564 (54.9%)	354 (12.3%)
<i>Food security: Yes</i>	3,017 (81.1%)	726 (71.0%)	2,610 (91.8%)
<i>Veteran status: Yes</i>	487 (12.9%)	141 (13.6%)	443 (15.4%)
<i>Health insurance</i>			
Uninsured	299 (8.0%)	50 (4.9%)	190 (6.7%)
Any private insurance	1,519 (40.8%)	319 (31.2%)	1,564 (55.0%)
Public insurance only	1,908 (51.2%)	655 (64.0%)	1,092 (38.4%)
<i>Active cancer: Yes</i>	58 (1.5%)	34 (3.3%)	24 (0.8%)
<i>Diabetes: Yes</i>	1,270 (33.6%)	396 (38.2%)	765 (26.5%)
<i>Lung disease: Yes</i>	685 (18.1%)	287 (27.7%)	227 (7.9%)
<i>Angina: Yes</i>	293 (8.3%)	127 (13.4%)	90 (3.2%)
<i>Stroke: Yes</i>	373 (9.9%)	134 (12.9%)	164 (5.7%)
<i>BMI category</i>			
Underweight (BMI < 18.5)	61 (1.6%)	28 (2.7%)	34 (1.2%)
Normal weight (18.5 ≤ BMI < 25)	751 (20.0%)	206 (20.0%)	685 (23.9%)
Overweight (25 ≤ BMI < 30)	1,177 (31.4%)	292 (28.3%)	1,058 (36.9%)

Variable	Class		
	1 High-Impact Pain (OTC) (n = 3,787)	2 High-Impact Severe Pain (Opioids) (n = 1,038)	3 Low-Impact Mild Pain (n = 2,887)
Obese 1 ($30 \leq \text{BMI} < 35$)	903 (24.1%)	238 (23.1%)	668 (23.3%)
Obese 2 ($35 \leq \text{BMI} < 40$)	484 (12.9%)	139 (13.5%)	263 (9.2%)
Obese 3 (BMI 40+)	373 (9.9%)	129 (12.5%)	162 (5.6%)
<i>Heart condition: Yes</i>	1,169 (30.9%)	378 (36.5%)	626 (21.7%)
<i>Arthritis: Yes</i>	2,996 (79.2%)	910 (87.8%)	1,946 (67.5%)
<i>Smoker status</i>			
Never smoker	1,507 (40.0%)	282 (27.3%)	1,332 (46.3%)
Former smoker	1,493 (39.6%)	452 (43.8%)	1,203 (41.8%)
Current smoker	771 (20.4%)	299 (28.9%)	340 (11.8%)
<i>Depressive symptoms (CESD score)</i>	2.0 (3.0)	3.0 (5.0)	1.0 (2.0)

All physical health conditions considered are also most prevalent in the **High-Impact Severe Pain (Opioids)** class, with 3.3% of class members reporting active cancer, 38.2% reporting diabetes, 27.7% reporting lung disease, 13.4% reporting angina, 12.9% reporting stroke, 36.5% reporting a heart condition, and 87.8% reporting arthritis. The **High-Impact Severe Pain (Opioids)** class also has the highest median CESD depressive symptoms score (3.0, IQR = 5.0), compared to a median of 2.0 (IQR = 3.0) in the **High-**

Impact Pain (OTC) class and a median of 1.0 (IQR = 2.0) in the **Low-Impact Mild Pain** class. Having a body mass index (BMI) categorised as “Normal weight” ($18.5 \leq \text{BMI} < 25$) is most common in the **Low-Impact Mild Pain** class (23.9%). Meanwhile, 20.0% of both the **High-Impact Pain (OTC)** class and the **High-Impact Pain (OTC)** class are composed of individuals with a “Normal weight” BMI. Never having smoked is also most common in the **Low-Impact Mild Pain** class (46.3%), while 40.0% of the **High-Impact Pain (OTC)** class and 27.3% of the **High-Impact Severe Pain (Opioids)** class have never smoked.

6.4 Discussion

This study identified three distinct latent classes of pain experience in a large sample of older American adults. These latent classes are intuitively interpretable in the context of pain experience, appearing to represent a low impact pain group and two higher impact pain groups that are distinguished by different pain medication use and reporting of severe pain.

The three identified pain experience latent classes are also distinguishable by the sociodemographic and health profiles of their members. The set of sociodemographic and health characteristics summarised to explore the composition of each latent class was adopted from the confounder adjustment set used in the Chapter 4 HRS analysis. Some of these factors (sex, age, education level, health insurance) were also examined in the Chapter 3 investigation of sociodemographic disparities in the TILDA dataset. The group considered to be most impacted by pain in this analysis [**High-Impact Severe Pain (Opioids)**] report lower levels of education and private health insurance than the latent classes less impacted by pain, as well as having the highest

proportion of female members. This pattern is similar to the results of the TILDA analysis in Chapter 3, where lower educational attainment, not having private health insurance, and female sex were all associated with more severe pain. Such sociodemographic disparities in pain experiences have also been reported in other countries (Milani et al., 2022; Stewart Williams et al., 2015; Wranger et al., 2016; Zimmer et al., 2020). These disparities warrant attention from policy makers aiming to tackle disproportionate societal pain burdens. Additionally, the **High-Impact Severe Pain (Opioids)** group were the most likely to have each of the other chronic health conditions considered in this analysis.

In addition to exploring a more holistic pain experience measure, this study demonstrates the issues of LCA model selection that were discussed previously in Chapter 5. The fit indices considered did not unanimously select the same number of latent classes as the optimal solution for the final model, making the task of model selection more challenging. While the BIC, NFI, and NNFI favoured the 3-class model, both the BLRT and the AIC suggested models with a larger number of latent classes. This tendency for the AIC to select a more complex model when sample size is large was observed previously in Nylund et al. (2007) and in Chapter 5 of this thesis. Additionally, the bivariate residuals (BVRs) for the 3-class model suggested there may be some residual local dependence between the indicator variables in the model. However, the 3-class model was well identified and provided a more interpretable solution compared to the candidate models with more latent classes, which are both desirable characteristics for an LCA model (Collins and Lanza, 2010). The NFI and NNFI interpreted using the elbow rule suggested in Chapter 5 proved valuable in helping to guide model selection in this analysis, as the new fit indices provided additional support for ultimately selecting the 3-class model

as the most parsimonious fit.

While BVRs are a useful tool for assessing the local independence assumption and selecting an appropriate number of latent classes for the final model, they are subject to limitations. BVRs are typically not reliable when there is missing data on the observed indicator variables used to fit the LCA model (Asparouhov and Muthén, 2015). Note, a complete cases analysis was conducted for this study, as only a small number of participants (4.6%) with pain in the 2016 HRS sample were missing data on at least one pain indicator variable. Another issue with BVRs is that they calculate the total residual association across latent classes rather than distinguishing class-specific residual associations (Asparouhov and Muthén, 2015). As a result, it is possible for a positive residual in one latent class to cancel out a negative residual in another latent class, in which case the total residual does not give an indication of class dependence.

The poLCA R package (Linzer and Lewis, 2011) used to fit the LCA models in this study also has some limitations. The package does not currently support the inclusion of sample weights in the analysis, thus HRS sample weights could not be applied to increase the generalisability of the results to the entire population of American older adults with pain in 2016. Additionally, while methods have been developed for modelling correlated residuals when the local independence assumption of LCA is not met (Visser and Depaoli, 2022), this methodology has not been incorporated into the poLCA package. Some residual correlations were detected between back pain and OTC medication use (BVR = 5.53, p-value = 0.01) and between back pain and opioid medication use (BVR = 4.64, p-value = 0.04) in the final 3-class LCA model. Modelling these local dependencies would likely improve the accuracy of the model parameter estimates. The authors of the poLCA package have indicated

further development to include both sample weighting and local dependence modelling in future versions of the package (Linzer and Lewis, 2011).

The study is also limited by the availability of different pain measures in the HRS dataset. For instance, while it would likely be beneficial to include information on other pain locations in the LCA model, there is a lack of data on pain location in the HRS dataset beyond back pain. Attitudes towards pain such as fear and avoidance have also been highlighted as important influencers of the pain experience (Sullivan and Ballantyne, 2016; Van Hecke et al., 2013), however such information was not collected from the 2016 HRS wave participants. An important aim for future work will be to identify pain experience latent classes using a broader, more diverse range of pain measures. Future work analysing older adult studies in other countries would also be beneficial to investigate the generalisability of the identified pain experience latent classes to other populations. Cultural and country-specific differences in reported pain prevalence (Pillay et al., 2014) and pain management (Breivik et al., 2006) have been observed previously, thus it is possible that the structure of pain experience latent classes would also differ between countries. The global networking of ageing studies of which the HRS is a member is a valuable potential data source for such future inter-country comparisons (Gateway to Global Aging, 2023).

Utilising the identified latent classes as a variable to represent pain experience will be another key direction for future work. For example, causes of pain experience latent class membership could be investigated to further understanding of the determinants of pain experience. Possible methodological approaches to such causal analyses with a latent class *outcome* have been outlined by Lanza et al. (2013) and (Clouth et al., 2022). Additionally, the causal effect of pain experience latent class on health or mortality outcomes

of interest could be analysed, similar to the investigation of the effect of pain intensity on mortality presented in Chapter 4. Various approaches to incorporating a latent class *exposure* in causal analyses have also been proposed, as detailed in Clouth et al. (2023).

Overall, this study demonstrates the suitability of LCA methodologies to identify latent classes of pain experience using a sample of American older adults. These latent classes give a more comprehensive representation of the individual's pain experience than single variable measures, by incorporating multiple dimensions of pain including pain intensity, pain location, pain-related disability, and pain medication use. The identification of such latent classes may be a key first step in future work investigating how various upstream factors influence pain experience, and also how pain experience impacts future health and mortality outcomes.

7 Conclusions and Future Research

7.1 Conclusions

As outlined in Chapter 1, the goal of this thesis was to address challenges around biases, causality and latent class methodologies when modelling pain in older adult cohort studies. The research included in this thesis has addressed this goal, beginning with an analysis of five waves of data from TILDA in Chapter 3. The aim of the publication presented in Chapter 3 was to model previously unexplored longitudinal sociodemographic disparities in pain in Irish older adults, and to investigate attrition bias, mortality bias, and reporting heterogeneity measurement bias that could compromise the accuracy of pain estimates. This study was the first to model pain over five waves of TILDA and adds to the body of knowledge on sociodemographic disparities in pain in older adults. Evidence of the sex and socioeconomic differences previously reported in other countries (Cimas et al., 2018; Grol-Prokopczyk, 2017; Ikeda et al., 2019; Lacey et al., 2013; Stewart Williams et al., 2015) was also found in the TILDA sample. Thus, this work contributes to the triangulation of evidence and provides support for the generalisability of these patterns of disparities to other older adult cohorts. No significant evidence of pain-related attrition bias was found in TILDA, however there was evidence of pain related mortality bias. This finding is similar to previous results

regarding attrition and mortality bias in the HRS (Grol-Prokopczyk, 2017). However, while some evidence of systematic socioeconomic differences in pain reporting styles (reporting heterogeneity) was found in Chapter 3, gender differences in reporting observed in the HRS cohort (Grol-Prokopczyk, 2017) were not observed in TILDA. This highlights the importance of both country-specific analyses and cross-country comparisons to further the understanding of similarities and differences in pain experience and pain reporting across international older adult populations.

The publication presented in Chapter 4 addressed another statistical modelling challenge in older adult pain research. The aim of Chapter 4 was to investigate the causal effect of pain on 20-year mortality hazard in American older adults. As pain exposure cannot be randomised to investigate causal effects using experimental studies for ethical reasons, the observational HRS was used as a data source for this analysis. A DAG was designed using existing expert knowledge of the pain-mortality relationship to aid the identification of an adequate confounder adjustment set, and to transparently present the assumptions underlying the analysis. Advanced statistical methods involving propensity scores were applied to adjust for confounding, with the aim of achieving conditional exchangeability of the pain exposure and no-pain exposure groups in the sample. The estimated hazard ratios (HRs) for the main analyses suggested that pain exposure modestly increased mortality hazard. However, the results were also compatible with no effect, as the 95% confidence intervals for the HRs narrowly contained 1.0. Upstream factors that raise both pain and mortality risk, such as body mass index and depressive symptoms, were identified as potentially modifiable factors for interventions aimed at tackling pain or mortality. Additionally, multiple sensitivity analyses were conducted to investigate the robustness of the main results. As investigating

potential attrition selection bias in the TILDA older adult cohort was a central topic of Chapter 3, the analyses in Chapter 4 were conducted with and without incorporating the HRS sample weights to adjust for potential attrition selection bias. The analyses with and without sample weights gave similar results, reflecting the lack of evidence of attrition bias found in Chapter 4 and in the HRS study mentioned previously (Grol-Prokopczyk, 2017). Given the difficulty in measuring and defining pain exposure, multiple sensitivity analyses with different pain exposure definitions (e.g., severe pain only; moderate/severe pain; moderate/severe pain AND arthritis) were also conducted to assess the robustness of the results to different pain exposure definitions. Overall, conclusions about the potential causal effect were broadly similar for all alternative pain exposure definitions. To the authors' knowledge, the publication presented in Chapter 4 is the first to apply propensity score methods to investigate the potentially causal relationship between pain and mortality. This work thus provides a novel procedure for future causal analyses involving a pain exposure.

Self-reported pain severity (none, mild, moderate, severe) formed the basis of the pain measures analysed in both Chapter 3 and Chapter 4. However, as highlighted in Chapter 3, pain severity as a single measure is likely subject to some measurement bias arising from sociodemographic differences in reporting styles (reporting heterogeneity). Further, pain severity captures just one dimension of pain and its impact on the individual. Identifying a more holistic measure of pain experience using LCA is of interest. Chapter 5 addresses methodological issues of model fit indices when using LCA in large cohort studies. Firstly, the adaptation of fit indices used to guide model selection in structural equation modelling is proposed for use with LCA. The goal of this work was to aid the selection of an appropriate number of classes

when fitting LCA models using a large sample, such as the HRS dataset. The proposed LCA fit indices were evaluated through the development and application of simulation studies which investigated the performance of the fit indices across different sample sizes and different underlying population models. These simulation studies suggested that “rule-of-thumb” cut-offs for interpretation may not be appropriate. However, applying an “elbow-rule” for interpretation could identify a more parsimonious solution when compared to some existing tools used for LCA class enumeration. While definitive guidelines could not be established for interpreting the proposed LCA fit indices, the potential of the elbow rule was demonstrated in the application detailed in Chapter 6. However, how exactly to identify the “elbow” in a plot is somewhat arbitrary and the proposed fit indices should be jointly considered along with other model selection criteria such as the BIC, identifiability of the maximum likelihood solution, and interpretability of the identified latent classes to select the most parsimonious solution.

The goal of Chapter 6 was to develop an LCA model to identify unique classes of pain experience in the 2016 HRS sample. The analysis succeeded in identifying a holistic measure of pain experience that incorporated multiple aspects of pain including severity, disability, location, and over-the-counter and opioid pain medication use. The fit indices proposed in Chapter 5 were used to assist in the selection of the final model, together with considering the BIC and the identifiability and interpretability of the candidate models. An examination of the sociodemographic profiles of each pain experience class also confirmed some of the pain disparities observed earlier in Chapter 3. Female sex and measures of poorer socioeconomic background such as lower educational attainment were most common in the class with highest pain impact and opioid use. Similar sex and socioeconomic disparities in pain have

also been observed in other studies (Milani et al., 2022; Stewart Williams et al., 2015; Wrangler et al., 2016; Zimmer et al., 2020). Additionally, all chronic conditions considered were most common in this highest pain impact class. Chapter 6 has laid important groundwork for further analyses using holistic LCA pain measures to capture pain experience. Additionally, the fit indices proposed in Chapter 5 proved helpful in contributing to the selection of the final 3-class LCA model discussed in Chapter 6. These fit indices may aid future LCA model selection both within pain research and in other disciplines.

The results presented in this thesis are subject to some limitations. Firstly, the self-reported nature of the TILDA and HRS pain measures used in this thesis is an important limitation of pain modelling using cohort studies. As discussed in Chapter 3, these measures are subjective and thus liable to be affected by measurement bias such as reporting heterogeneity. Reducing reliance on a single pain measure by instead using a composite pain measure like the latent class variable identified in Chapter 6 may reduce the influence of such measurement bias in individual measures, however it will not remove the bias completely.

The range of different pain measures available in the HRS dataset was also a limitation of the analysis in Chapter 6. Data on some dimensions of pain, such as attitudes towards pain (Sullivan and Ballantyne, 2016; Van Hecke et al., 2013), was not available in the HRS and thus could not be included as indicator variables in the LCA model. Additionally, the `poLCA` R package used to fit the LCA models in Chapter 6 has some limitations. The package does not have capabilities to incorporate sample weighting when fitting LCA models, and it is also not currently possible to model local dependences (Linzer and Lewis, 2011).

Finally, the counterfactual approach to causal inference using observational data applied in Chapter 4 has faced some criticisms. The focal limitation of this approach is the unverifiable assumption that all confounders have been adjusted for in the analysis, i.e., that there is no unmeasured confounding. In addition to requiring sufficient data to measure all confounders, this assumption is also contingent on correctly identifying an adequate confounder adjustment set. However, determining the causal structure underlying the causal relationship of interest is a difficult task, particularly when expert knowledge does not agree. This challenge was demonstrated in Chapter 4, where there was disagreement in the existing literature regarding whether depressive symptoms should be treated as a confounder or a mediator of the pain-mortality causal relationship. Possible approaches to address some of these limitations are outlined in the next section, along with other directions for future research.

7.2 Future research

The cohorts studied in this thesis, TILDA and the HRS, are part of a global network of older adult cohort studies that could be exploited in future pain research building on the methods and results presented in this thesis. The global network of studies spans over 30 other countries, including China (CHARLS; Zhao et al., 2014), Brazil (ELSI-Brazil; Lima-Costa et al., 2023), and England (ELSA; Steptoe et al., 2013), with the full list of international studies included on the Gateway to Global Aging Data website: <https://g2aging.org/survey/overview>. The similar study designs and often identical phrasing of questions used across studies facilitates the direct comparison of findings between countries. Such cross-country comparison

could examine the generalisability of the results around sociodemographic disparities and potential biases presented in Chapter 3 and the pain experience latent classes identified in Chapter 6. Additionally, replicating the pain-mortality causal analysis presented in Chapter 4 using older adult cohorts from different countries would serve as a form of triangulation of evidence (Krieger and Davey Smith, 2016) for the causal results reported therein.

While this global collection of similar ageing studies is a valuable data source for future investigations of causal relationships in older adult pain research, care must be taken to adapt each analysis to the specific healthcare and social contexts of each country. As discussed in Chapter 4, identifying the causal structure likely to underlie the causal relationship under study is a non-trivial task requiring careful consideration and expert knowledge. Thoughtful DAG design for cohort studies is important not only to aid confounder identification, but to prevent inducing bias by over-adjusting for mediator or collider variables (Van Zwieten et al., 2022; VanderWeele, 2009b). Ferguson et al. (2020) proposed a systematic approach to the task of evidence synthesis to inform DAG creation, which may be useful to guide future research.

Investigating the potential cumulative causal effect of pain at multiple time points on mortality risk will be another important direction for future work. Such longitudinal analyses are made more complex by treatment-confounder feedback loops, which occur if a confounder affects the treatment (or exposure) and the treatment (or exposure) also affects that confounder at a later time point (Hernán and Robins, 2020), such as the bidirectional relationship between pain exposure and depressive symptoms discussed in Chapter 4. In the presence of treatment-confounder feedback, the confounder adjustment methods for fixed treatments/exposures applied in Chapter 4 (propensity score matching, inverse probability weighting) are unable to account for confounder

bias without introducing additional biases (Robins and Hernán, 2008). Future work could extend this analysis to investigate the effects of time-varying pain exposure on mortality by developing statistical models based on g-methods (Naimi et al., 2017).

Further research could examine the extent of reporting heterogeneity in self-reported pain severity (and in other self-reported pain measures), potentially incorporating objective health measures that may be correlated with pain experience such as walking speed (Hicks et al., 2017; Simmonds et al., 2012) or a frailty index (Calvey et al., 2022). Expanding on the development of pain experience latent classes in Chapter 6 is another direction for future work on the topic of pain measurement. Replicating the analysis using other data sources and sets of pain indicator variables, ideally from different populations, will also serve to assess the generalisability of the identified classes. Additionally, the authors of the `poLCA` R package have indicated that both sample weighting and local dependence modelling capabilities will be included in future versions of the package (Linzer and Lewis, 2011). It will be of interest to apply these capabilities in the future to further the modelling of pain experience latent classes. The LCA fit indices proposed in Chapter 5 could be used to aid model selection when identifying these pain experience latent classes, to further understanding of their behaviour and utility. Additional simulation studies and real data applications will also be beneficial to further explore the utility and interpretation of these fit indices.

Another interesting direction for future research could be to extend the causal statistical models of pain on survival developed in Chapter 4 to incorporate latent pain classes (such as those developed in Chapter 6) as the exposure of interest. While approaches for fitting an LCA model with a survival distal outcome have been established (Lythgoe et al., 2019), to

the author's knowledge there are no existing studies in the literature that combine a causal framework with LCA to model the effect of a latent class exposure on a survival distal outcome. As individuals cannot be randomised to latent exposure classes due to the unobservable nature of the classes, various confounder adjustment approaches have been proposed to aid estimation of the causal effect of a latent class exposure (Bray et al., 2019; Schuler et al., 2014; Yamaguchi, 2015). Clouth et al. (2023) recently introduced two new bias-adjusted three-step approaches which aim to combine the strengths of these earlier methods. However, attempts to conduct causal analyses with latent variable exposures have received some criticism, as outlined in Clouth et al. (2023). VanderWeele (2022) argues that the potential outcomes under a latent variable exposure are ill-defined, and so a latent variable exposure may constitute a breach of the consistency assumption for causal inference (VanderWeele, 2009a). VanderWeele (2022) has proposed to tackle this issue with a new measurement model for latent exposures (VanderWeele and Hernán, 2013), however how to appropriately implement the LCA exposure modelling approaches developed by Clouth et al. (2023) within the framework suggested by VanderWeele (2022) is also a question for future research.

In conclusion, this thesis tackled multiple important challenges in the modelling of pain using older adult cohort studies. Potential biases, causal inference, and latent class methodologies were among the key topics addressed. All findings were discussed within the context of the literature and limitations of the approaches taken. Directions for future research to expand upon the work presented in this thesis were also outlined.

A Supplementary Material for Chapter 5

This appendix contains additional results from the simulation studies conducted to investigate the behaviour of the LCA fit indices detailed in Chapter 5. Section A.1 below pertains to simulation study 1, and includes summary figures and tables for the simulations with *unequal* class sizes. These summaries mirror the results for the simulations with *equal* class sizes that are included in Subsection 5.5.1 of the main text. Section A.2 of this appendix contains a figure and table summarising the behaviour of the proposed NNFI fit statistic for the models fitted in simulation study 2.

A.1 Simulation study 1: Results for unequal class sizes

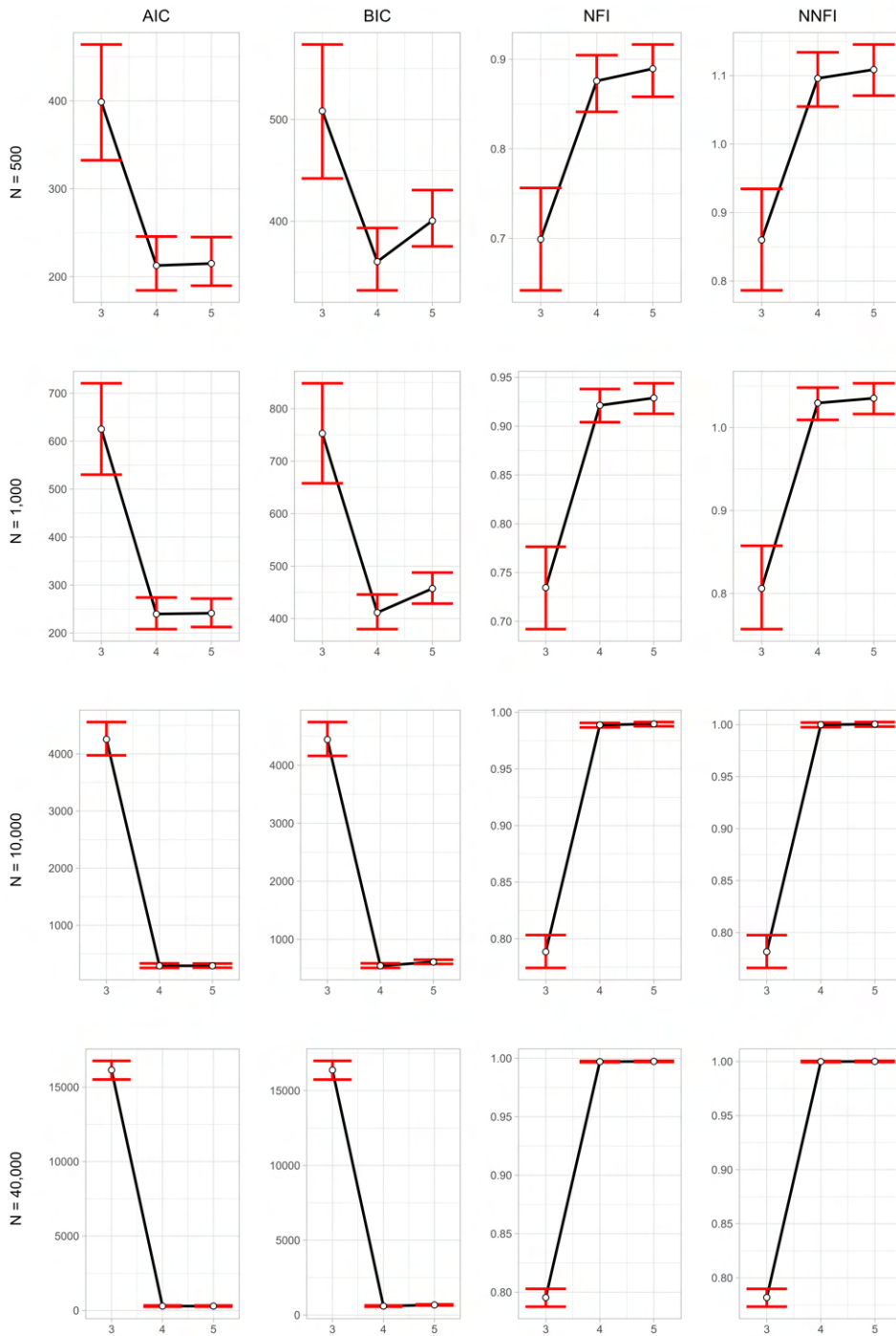


Figure A.1: Distributions of fit indices values for the **structure A strong pattern** simulations with **unequal class sizes**. *Note:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

Table A.1: Percentage of **structure A strong pattern** simulated data sets with **unequal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	0.0%	76.7%	23.3%
	AIC within 2	0.0%	88.5%	11.5%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	5.4%	94.6%
	NNFI >.95	1.1%	98.9%	0.0%
<i>N=1,000</i>	Lowest AIC	0.0%	70.5%	29.5%
	AIC within 2	0.0%	84.3%	15.7%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	99.1%	0.9%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=10,000</i>	Lowest AIC	0.0%	59.2%	40.8%
	AIC within 2	0.0%	76.2%	23.8%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=40,000</i>	Lowest AIC	0.0%	58.6%	41.4%
	AIC within 2	0.0%	74.7%	25.3%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

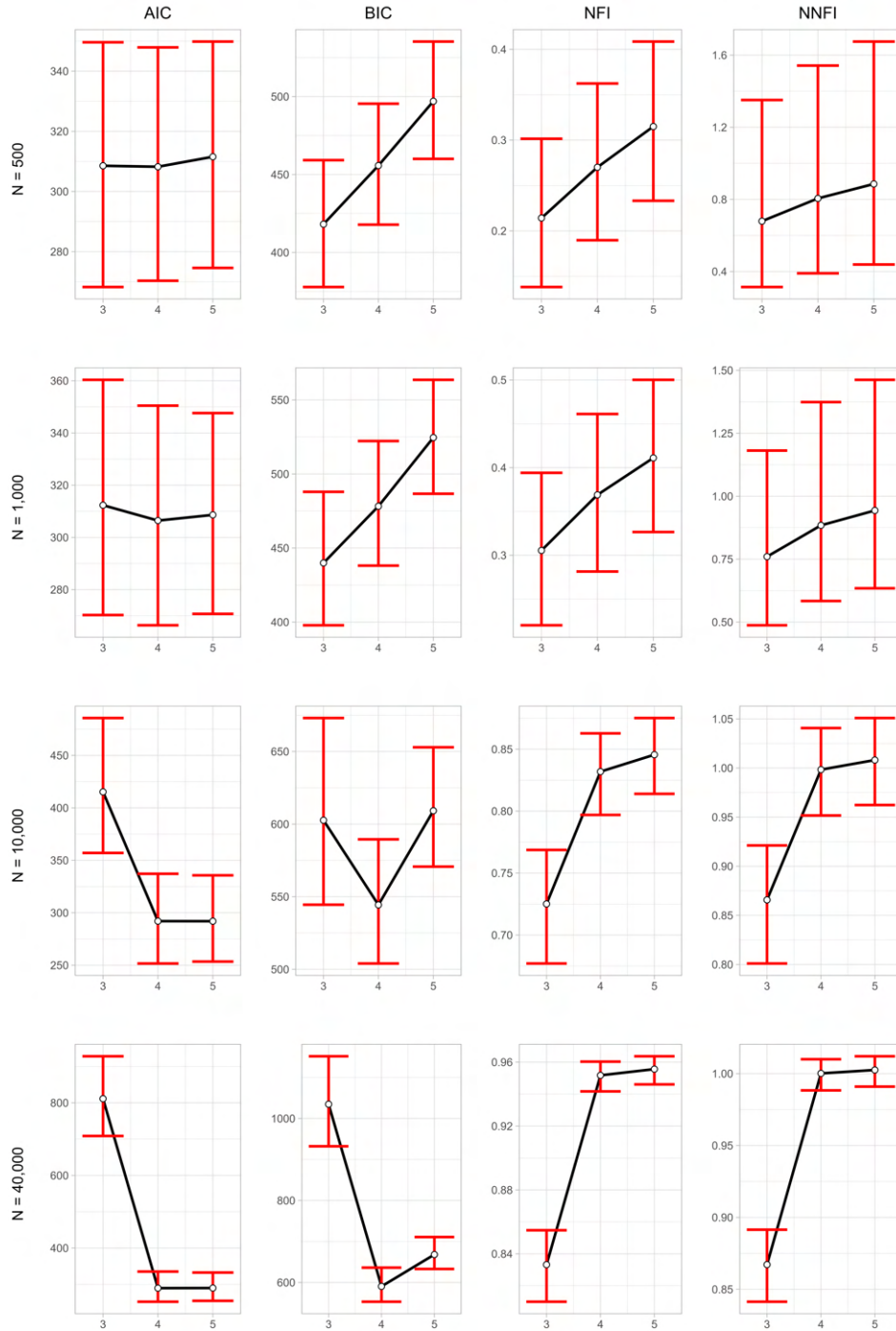


Figure A.2: Distributions of fit indices values for the **structure A weak pattern** simulations with **unequal class sizes**. *Note:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

Table A.2: Percentage of **structure A weak pattern** simulated data sets with **unequal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	51.9%	36.4%	11.7%
	AIC within 2	68.9%	26.6%	4.5%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	10.2%	11.2%	78.6%
<i>N=1,000</i>	Lowest AIC	20.0%	56.3%	23.7%
	AIC within 2	29.9%	58.3%	11.8%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	10.0%	20.3%	69.7%
<i>N=10,000</i>	Lowest AIC	0.0%	54.9%	45.1%
	AIC within 2	0.0%	73.2%	26.8%
	Lowest BIC	0.3%	99.7%	0.0%
	BIC within 2	0.4%	99.6%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	0.2%	97.7%	2.1%
<i>N=40,000</i>	Lowest AIC	0.0%	58.4%	41.6%
	AIC within 2	0.0%	73.6%	26.4%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

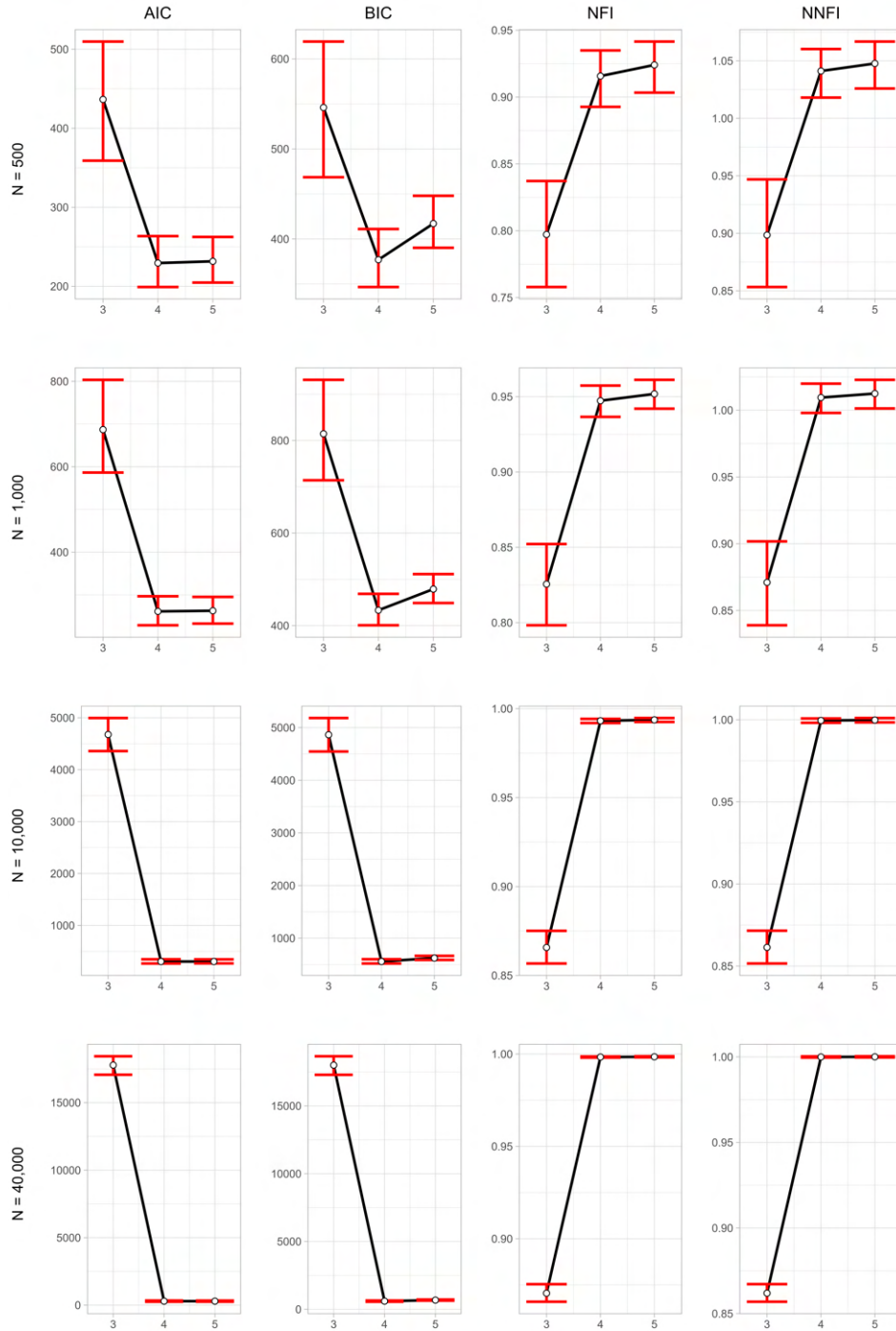


Figure A.3: Distributions of fit indices values for the **structure B strong pattern** simulations with **unequal class sizes**. *Note:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

Table A.3: Percentage of **structure B strong pattern** simulated data sets with **unequal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	0.0%	76.5%	23.5%
	AIC within 2	0.0%	87.9%	12.1%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	93.3%	6.7%
	NNFI >.95	1.5%	98.5%	0.0%
<i>N=1,000</i>	Lowest AIC	0.0%	68.7%	31.3%
	AIC within 2	0.0%	83.2%	16.8%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=10,000</i>	Lowest AIC	0.0%	53.9%	46.1%
	AIC within 2	0.0%	73.2%	26.8%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%
<i>N=40,000</i>	Lowest AIC	0.0%	52.8%	47.2%
	AIC within 2	0.0%	70.5%	29.5%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	0.0%	100.0%	0.0%
	NNFI >.95	0.0%	100.0%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

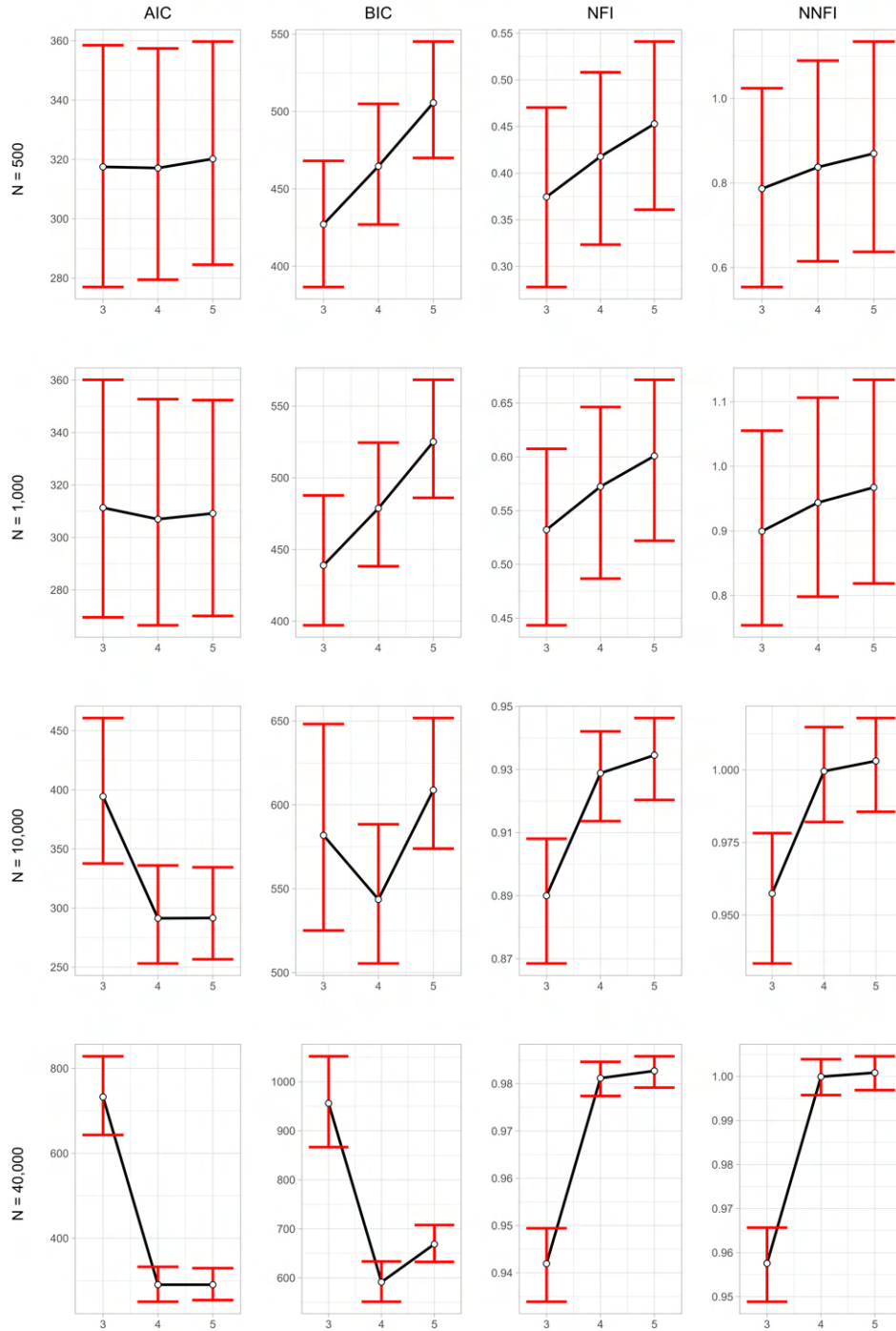


Figure A.4: Distributions of fit indices values for the **structure B weak pattern** simulations with **unequal class sizes**. *Note:* For each plot, the x-axis includes 3-5 classes; the y-axis is the fitted value of the given model selection tool; each row of plots corresponds to a different sample size. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

Table A.4: Percentage of **structure B weak pattern** simulated data sets with **unequal class sizes** in which each number of latent classes are selected by each model selection tool.

Sample Size	Tool	Number of Classes		
		3	4	5
<i>N=500</i>	Lowest AIC	49.2%	36.6%	14.2%
	AIC within 2	65.1%	29.6%	5.3%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	0.0%	8.1%	83.9%
<i>N=1,000</i>	Lowest AIC	23.6%	53.5%	22.9%
	AIC within 2	36.9%	53.2%	9.9%
	Lowest BIC	100.0%	0.0%	0.0%
	BIC within 2	100.0%	0.0%	0.0%
	NFI >.90	0.0%	0.0%	100.0%
	NNFI >.95	24.6%	20.5%	54.9%
<i>N=10,000</i>	Lowest AIC	0.0%	57.5%	42.5%
	AIC within 2	0.0%	74.8%	25.2%
	Lowest BIC	3.2%	96.8%	0.0%
	BIC within 2	4.8%	95.2%	0.0%
	NFI >.90	15.4%	84.6%	0.0%
	NNFI >.95	73.8%	26.2%	0.0%
<i>N=40,000</i>	Lowest AIC	0.0%	55.3%	44.7%
	AIC within 2	0.0%	72.4%	27.6%
	Lowest BIC	0.0%	100.0%	0.0%
	BIC within 2	0.0%	100.0%	0.0%
	NFI >.90	100.0%	0.0%	0.0%
	NNFI >.95	95.3%	4.7%	0.0%

Notes: ‘AIC within 2’ and ‘BIC within 2’ use a decision rule that adopts the most parsimonious model within 2 units of best AIC or BIC, respectively. As only models with 3-5 classes were considered, the 3-class and 5-class columns denote the percentage of simulations for which the fit indices suggest ≤ 3 or ≥ 5 classes respectively.

A.2 Simulation study 2: NNFI results

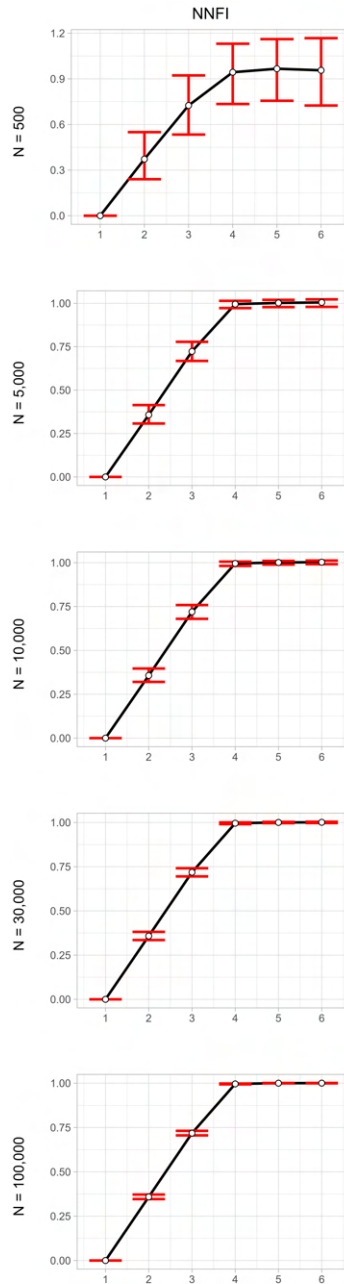


Figure A.5: Distributions of NNFI values for models with 1-6 classes for different sample sizes. *Notes:* The x-axis includes 1-6 classes; the y-axis is the fitted value of the given model selection tool. The true number of classes specified by the simulation design is 5, but 4 classes are considered adequate for a practically reasonable answer. The empty circles represent means across all replication samples and the red lines represent the span that covers 95% of the values across replication samples.

Table A.5: Percentage of simulated data sets in which latent class model sizes are selected by the NNFI selection rules.

Sample Size	Tool	Number of Classes					
		1	2	3	4	5	6
<i>N=500</i>	NNFI “elbow”	0.0%	4.1%	32.8%	61.7%	1.4%	0.0%
	NNFI >.95	0.0%	0.0%	1.6%	46.1%	11.6%	40.7%
<i>N=5,000</i>	NNFI “elbow”	0.0%	0.0%	1.2%	98.8%	0.0%	0.0%
	NNFI >.95	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
<i>N=10,000</i>	NNFI “elbow”	0.0%	0.0%	0.1%	99.9%	0.0%	0.0%
	NNFI >.95	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
<i>N=30,000</i>	NNFI “elbow”	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	NNFI >.95	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
<i>N=100,000</i>	NNFI “elbow”	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	NNFI >.95	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%

B Supplementary Material for Chapter 6

B.1 Summary of background characteristics for those with and without missing pain data

Table B.1: Comparison of background characteristics between the sample with full pain data and the participants removed from the sample due to missing pain data.

Variable	No missing pain data (n = 7,712)	Missing pain data (n = 373)	P-value
<i>Age in years</i>	64.0 (57.0, 75.0)	70.0 (60.0, 80.0)	<0.001
<i>Sex: Female</i>	4,768 (61.8%)	267 (71.6%)	<0.001
<i>Race/ethnicity</i>			0.223
White (non-Hispanic)	4,424 (57.4%)	206 (55.2%)	
Black (non-Hispanic)	1,658 (21.5%)	77 (20.6%)	
Hispanic	1,226 (15.9%)	74 (19.8%)	
Other (non-Hispanic)	399 (5.2%)	16 (4.3%)	
<i>Marital status</i>			<0.001

Variable	No missing pain data (n = 7,712)	Missing pain data (n = 373)	P-value
Married	3,870 (50.3%)	151 (40.7%)	
Separated/divorced	1,773 (23.0%)	87 (23.5%)	
Widowed	1,395 (18.1%)	107 (28.8%)	
Never married	638 (8.3%)	25 (6.7%)	
Other	24 (0.3%)	1 (0.3%)	
<i>Household size</i>	2.0 (2.0, 3.0)	2.0 (1.0, 3.0)	0.009
<i>Number of children</i>	3.0 (2.0, 4.0)	3.0 (2.0, 4.0)	0.611
<i>Region</i>			0.747
Northeast	1,108 (14.4%)	48 (12.9%)	
Mid-west	1,547 (20.1%)	71 (19.1%)	
South	3,431 (44.6%)	175 (47.2%)	
West	1,610 (20.9%)	77 (20.8%)	
<i>Urbanicity</i>			0.440
Urban	4,038 (52.5%)	182 (49.2%)	
Suburban	1,551 (20.2%)	78 (21.1%)	
Ex-urban/rural	2,100 (27.3%)	110 (29.7%)	
<i>Education level</i>			<0.001
No degree	1,538 (21.1%)	112 (31.0%)	
High school degree	4,323 (59.3%)	205 (56.8%)	
4-year college degree	862 (11.8%)	28 (7.8%)	
Graduate degree	561 (7.7%)	16 (4.4%)	
<i>Wealth quartile</i>			0.405
Q1	2,450 (31.8%)	128 (34.3%)	
Q2	2,044 (26.5%)	106 (28.4%)	

Variable	No missing pain data (n = 7,712)	Missing pain data (n = 373)	P-value
Q3	1,702 (22.1%)	74 (19.8%)	
Q4 (wealthiest)	1,515 (19.6%)	65 (17.4%)	
<i>Employment status</i>			<0.001
Employed	2,315 (30.3%)	71 (19.0%)	
Unemployed	302 (3.9%)	10 (2.7%)	
Retired	2,896 (37.8%)	165 (44.2%)	
Not in labour force	2,139 (28.0%)	127 (34.0%)	
<i>Food security: Yes</i>	6,353 (83.7%)	301 (82.2%)	0.494
<i>Veteran status: Yes</i>	1,071 (13.9%)	42 (11.3%)	0.178
<i>Health insurance</i>			0.002
Uninsured	539 (7.1%)	15 (4.2%)	
Any private insurance	3,402 (44.8%)	137 (38.5%)	
Public insurance only	3,655 (48.1%)	204 (57.3%)	
<i>Active cancer: Yes</i>	116 (1.5%)	8 (2.2%)	0.435
<i>Diabetes: Yes</i>	2,431 (31.5%)	137 (36.7%)	0.041
<i>Lung disease: Yes</i>	1,199 (15.6%)	71 (19.2%)	0.069
<i>Angina: Yes</i>	510 (7.0%)	30 (8.8%)	0.268
<i>Stroke: Yes</i>	671 (8.7%)	39 (10.5%)	0.277
<i>BMI category</i>			0.032
Underweight (BMI < 18.5)	123 (1.6%)	11 (3.0%)	
Normal weight (18.5 ≤ BMI < 25)	1,642 (21.5%)	98 (27.0%)	

Variable	No missing pain data (n = 7,712)	Missing pain data (n = 373)	P-value
Overweight ($25 \leq \text{BMI} < 30$)	2,527 (33.0%)	117 (32.2%)	
Obese 1 ($30 \leq \text{BMI} < 35$)	1,809 (23.6%)	74 (20.4%)	
Obese 2 ($35 \leq \text{BMI} < 40$)	886 (11.6%)	37 (10.2%)	
Obese 3 (BMI 40+)	664 (8.7%)	26 (7.2%)	
<i>Heart condition: Yes</i>	2,173 (28.2%)	118 (31.7%)	0.159
<i>Arthritis: Yes</i>	5,852 (76.0%)	303 (81.7%)	0.014
<i>Smoker status</i>			0.434
Never smoker	3,121 (40.6%)	151 (40.7%)	
Former smoker	3,148 (41.0%)	161 (43.4%)	
Current smoker	1,410 (18.4%)	59 (15.9%)	
<i>Depressive symptoms (CESD score)</i>	1.0 (0.0, 4.0)	2.0 (1.0, 4.0)	0.123

Note: P-values are reported for χ^2 tests for categorical variables or Kruskal-Wallis tests for numeric variables.

B.2 Bivariate residual summaries for the 2-class, 4-class, 5-class, and 6-class candidate LCA models

Table B.2: Bivariate residuals for the candidate 2-class LCA model.

Variables	Pain severity	Pain-related disability	OTC medication	Opioid medication
Pain-related disability	5.16 (0.00)			
OTC medication	9.43 (0.00)	3.33 (0.00)		
Opioid medication	11.85 (0.00)	2.38 (0.00)	39.16 (0.00)	
Back pain	1.19 (0.10)	8.25 (0.00)	5.62 (0.01)	8.62 (0.00)

Notes: Bootstrap p-values are included in brackets. OTC = over-the-counter.

Table B.3: Bivariate residuals for the candidate 4-class LCA model.

Variables	Pain severity	Pain-related disability	OTC medication	Opioid medication
Pain-related disability	0.04 (1.00)			
OTC medication	0.13 (1.00)	0.00 (1.00)		
Opioid medication	0.06 (1.00)	0.08 (1.00)	0.00 (1.00)	
Back pain	0.10 (0.99)	0.00 (1.00)	2.18 (0.30)	1.20 (0.84)

Notes: Bootstrap p-values are included in brackets. OTC = over-the-counter.

Table B.4: Bivariate residuals for the candidate 5-class LCA model.

Variables	Pain severity	Pain-related disability	OTC medication	Opioid medication
Pain-related disability	0.01 (1.00)			
OTC medication	0.01 (1.00)	0.00 (1.00)		
Opioid medication	0.12 (1.00)	0.02 (1.00)	0.02 (1.00)	
Back pain	0.67 (0.97)	0.03 (1.00)	0.29 (0.93)	0.05 (1.00)

Notes: Bootstrap p-values are included in brackets. OTC = over-the-counter.

Table B.5: Bivariate residuals for the candidate 6-class LCA model.

Variables	Pain severity	Pain-related disability	OTC medication	Opioid medication
Pain-related disability	0.01 (1.00)			
OTC medication	0.02 (1.00)	0.01 (0.99)		
Opioid medication	0.00 (1.00)	0.00 (1.00)	0.02 (1.00)	
Back pain	0.01 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)

Notes: Bootstrap p-values are included in brackets. OTC = over-the-counter.

Bibliography

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In F. BN Petrov (Ed.), *Proceedings of the Second International Symposium on Information Theory*, Budapest, Hungary, pp. 267–281. Akademiai Kiado.
- Alboukadel, K. (2022). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.5.0.
- Ali, M. S., R. H. Groenwold, W. R. Pestman, S. V. Belitser, K. C. Roes, A. W. Hoes, A. de Boer, and O. H. Klungel (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology and Drug Safety* 23(8), 802–811.
- Amoah, J., E. A. Stuart, S. E. Cosgrove, A. D. Harris, J. H. Han, E. Lautenbach, and P. D. Tamma (2020). Comparing propensity score methods versus traditional regression analysis for the evaluation of observational data: a case study evaluating the treatment of gram-negative bloodstream infections. *Clinical Infectious Diseases* 71(9), e497–e505.
- Andersson, H. I. (2009). Increased mortality among individuals with chronic widespread pain relates to lifestyle factors: a prospective population-based study. *Disability and Rehabilitation* 31(24), 1980–1987.
- Asparouhov, T. and B. Muthén (2015). Residual associations in latent class and latent transition analysis. *Structural Equation Modeling: A Multidisciplinary Journal* 22(2), 169–177.
- Atkinson, J. H., M. A. Slater, T. L. Patterson, I. Grant, and S. R. Garfin (1991). Prevalence, onset, and risk of psychiatric disorders in men with chronic low back pain: a controlled study. *PAIN[®]* 45(2), 111–121.
- Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3.
- Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 28(25), 3083–3107.
- Austin, P. C. (2009b). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics - Simulation and Computation* 38(6), 1228–1234.

- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46(3), 399–424.
- Austin, P. C. (2014a). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine* 33(6), 1057–1069.
- Austin, P. C. (2014b). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine* 33(7), 1242–1258.
- Austin, P. C. (2017). Double propensity-score adjustment: a solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research* 26(1), 201–222.
- Austin, P. C. and E. A. Stuart (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* 34(28), 3661–3679.
- Bago d’Uva, T., E. Van Doorslaer, M. Lindeboom, and O. O’Donnell (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics* 17(3), 351–375.
- Bair, M. J., R. L. Robinson, W. Katon, and K. Kroenke (2003). Depression and pain comorbidity: a literature review. *Archives of Internal Medicine* 163(20), 2433–2445.
- Banks, J., A. Kapteyn, J. P. Smith, and A. Van Soest (2009). *Work disability is a pain in the****, especially in England, the Netherlands, and the United States*, pp. 251–293. University of Chicago Press.
- Barrett, A., H. Burke, H. Cronin, A. Hickey, Y. Kamiya, R. A. Kenny, R. Layte, S. Maty, H. McGee, K. Morgan, et al. (2011). Fifty plus in Ireland 2011: first results from the Irish Longitudinal Study on Ageing (TILDA). *Dublin: Trinity College Dublin* 10, 2011–00.
- Bell, T., C. E. Franz, and W. S. Kremen (2022). Persistence of pain and cognitive impairment in older adults. *Journal of the American Geriatrics Society* 70(2), 449–458.
- Bentler, P. M. and D. G. Bonett (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88(3), 588.
- Bergmann, M., T. Kneip, G. De Luca, and A. Scherpenzeel (2019). Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1-7. Working Paper 41-2019, Survey of Health, Ageing and Retirement in Europe, Munich.

- Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* 2(3), 47–53.
- Biele, G., K. Gustavson, N. O. Czajkowski, R. M. Nilsen, T. Reichborn-Kjennerud, P. M. Magnus, C. Stoltenberg, and H. Aase (2019). Bias from self selection and loss to follow-up in prospective cohort studies. *European Journal of Epidemiology* 34, 927–938.
- Blyth, F. M., L. M. March, M. K. Nicholas, and M. J. Cousins (2003). Chronic pain, work performance and litigation. *PAIN[®]* 103(1-2), 41–47.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology* 53(1), 605–634.
- Bollen, K. A. and R. H. Hoyle (2012). Latent variables in structural equation modeling. *Handbook of Structural Equation Modeling* 1, 56–67.
- Bondesson, E., F. Larrosa Pardo, K. Stigmar, Å. Ringqvist, I. Petersson, A. Jöud, and M. Schelin (2018). Comorbidity between pain and mental illness—evidence of a bidirectional relationship. *European Journal of Pain* 22(7), 1304–1311.
- Bonevski, B., M. Randell, C. Paul, K. Chapman, L. Twyman, J. Bryant, I. Brozek, and C. Hughes (2014). Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology* 14, 1–29.
- Bowen, N. K. and S. Guo (2011). *Structural equation modeling*. Oxford University Press.
- Bradford Hill, A. (1965). The environment and disease: Association or causation? In *Proceedings of the Royal Society of Medicine*. Sage Publications.
- Brakenhoff, T. B., M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, and M. Van Smeden (2018). Measurement error is often neglected in medical literature: a systematic review. *Journal of Clinical Epidemiology* 98, 89–97.
- Bray, B. C., J. J. Dziak, M. E. Patrick, and S. T. Lanza (2019). Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later. *Prevention Science* 20, 394–406.
- Brayne, C. and T. E. Moffitt (2022). The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging. *Nature Aging* 2(9), 775–783.
- Breivik, H., B. Collett, V. Ventafridda, R. Cohen, and D. Gallacher (2006). Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *European Journal of Pain* 10(4), 287–333.

- Breivik, H., E. Eisenberg, and T. O'Brien (2013). The individual and societal burden of chronic pain in Europe: the case for strategic prioritisation and action to improve knowledge and availability of appropriate care. *BMC Public Health* 13, 1–14.
- Brunson, J. C. and Q. D. Read (2023). ggalluvial: Alluvial Plots in 'ggplot2'. R package version 0.12.5.
- Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33(2), 261–304.
- Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22(1), 31–72.
- Calvey, B., J. McHugh Power, and R. Maguire (2022). Expecting the best or fearing the worst: Discrepancies between self-rated health and frailty in an ageing Irish population. *British Journal of Health Psychology* 27(3), 971–989.
- Carreras, G., G. Miccinesi, A. Wilcock, N. Preston, D. Nieboer, L. Deliens, M. Groenvold, U. Lunder, A. Van Der Heide, and M. Baccini (2021). Missing not at random in end of life care studies: multiple imputation and sensitivity analysis on data from the ACTION study. *BMC Medical Research Methodology* 21(1), 1–12.
- Carroll, L. J., J. D. Cassidy, and P. Côté (2004). Depression as a risk factor for onset of an episode of troublesome neck and low back pain. *PAIN[®]* 107(1-2), 134–139.
- Chan, A., C. Malhotra, Y. K. Do, R. Malhotra, and T. Østbye (2011). Self reported pain severity among multiethnic older Singaporeans: Does adjusting for reporting heterogeneity matter? *European Journal of Pain* 15(10), 1094–1099.
- Chen, L., M. L. Ferreira, N. Nassar, D. B. Preen, J. L. Hopper, S. Li, M. Bui, P. R. Beckenkamp, B. Shi, N. K. Arden, et al. (2021). Association of chronic musculoskeletal pain with mortality among UK adults: A population-based cohort study with mediation analysis. *EClinicalMedicine* 42, 101202.
- Christensen, K., G. Doblhammer, R. Rau, and J. W. Vaupel (2009). Ageing populations: the challenges ahead. *The Lancet* 374(9696), 1196–1208.
- Cimas, M., A. Ayala, B. Sanz, M. S. Agulló-Tomás, A. Escobar, and M. Forjaz (2018). Chronic musculoskeletal pain in European older adults: Cross-national and gender differences. *European Journal of Pain* 22(2), 333–345.

- Cinelli, C. and C. Hazlett (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82(1), 39–67.
- Clouth, F., S. Pauws, F. Mols, and J. Vermunt (2022). A new three-step method for using inverse propensity weighting with latent class analysis. *Advances in Data Analysis and Classification* 16(2), 351–371.
- Clouth, F. J., T. T. Lê, and J. Vermunt (2023). Causal Latent Class Analysis with Distal Outcomes: A Modified Three-Step Method Using Inverse Propensity Weighting. Preprint: <https://osf.io/preprints/psyarxiv/tnea8> (Accessed: June 2024).
- Cohen, S. P., L. Vase, and W. M. Hooten (2021). Chronic pain: an update on burden, best practices, and new advances. *The Lancet* 397(10289), 2082–2097.
- Cole, S. R. and C. E. Frangakis (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 20(1), 3–5.
- Collins, L. M., P. L. Fidler, S. E. Wugalter, and J. D. Long (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research* 28(3), 375–389.
- Collins, L. M. and S. T. Lanza (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (First ed.). Hoboken, NJ: John Wiley & Sons.
- Covinsky, K. E., K. Lindquist, D. D. Dunlop, and E. Yelin (2009). Pain, functional limitations, and aging. *Journal of the American Geriatrics Society* 57(9), 1556–1561.
- Crimmins, E. M. and H. Beltrán-Sánchez (2011). Mortality and morbidity trends: is there compression of morbidity? *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 66(1), 75–86.
- Crombez, G., E. Veirman, D. Van Ryckeghem, W. Scott, and A. De Paepe (2023). The effect of psychological factors on pain outcomes: lessons learned for the next generation of research. *Pain Reports* 8(6), e1112.
- Curran, P. J., K. Obeidat, and D. Losardo (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development* 11(2), 121–136.
- Currie, S. R. and J. Wang (2005). More data on major depression as an antecedent risk factor for first onset of chronic back pain. *Psychological Medicine* 35(9), 1275–1282.

- Dahlhamer, J. (2018). Prevalence of chronic pain and high-impact chronic pain among adults—United States, 2016. *Morbidity and Mortality Weekly Report* 67(36), 1001–1006.
- Danaei, G., L. A. G. Rodríguez, O. F. Cantero, R. Logan, and M. A. Hernán (2013). Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical Methods in Medical Research* 22(1), 70–96.
- Danaei, G., M. Tavakkoli, and M. A. Hernán (2012). Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. *American Journal of Epidemiology* 175(4), 250–262.
- Davies, N. M., M. Dickson, G. Davey Smith, G. J. Van Den Berg, and F. Windmeijer (2018). The causal effects of education on health outcomes in the UK Biobank. *Nature Human Behaviour* 2(2), 117–125.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1), 1–22.
- Dennison, C. R. (2019). The crime-reducing benefits of a college degree: evidence from a nationally representative us sample. *Criminal Justice Studies* 32(4), 297–316.
- Deyo, R. A., S. F. Dworkin, D. Amtmann, G. Andersson, D. Borenstein, E. Carragee, J. Carrino, R. Chou, K. Cook, A. DeLitto, et al. (2015). Report of the NIH Task Force on research standards for chronic low back pain. *Physical Therapy* 95(2), e1–e18.
- Dezutter, J., L. A. Robertson, K. Luyckx, and D. Hutsebaut (2010). Life satisfaction in chronic pain patients: the stress-buffering role of the centrality of religion. *Journal for the Scientific Study of Religion* 49(3), 507–516.
- Dickerman, B. A., X. García-Albéniz, R. W. Logan, S. Denaxas, and M. A. Hernán (2019). Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine* 25(10), 1601–1606.
- Digitale, J. C., J. N. Martin, and M. M. Glymour (2022). Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology* 142, 264–267.
- Domenichiello, A. F. and C. E. Ramsden (2019). The silent epidemic of chronic pain in older adults. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 93, 284–290.

- Donoghue, O., M. Foley, and R. A. Kenny (2017). Cohort maintenance strategies used by the Irish Longitudinal Study on Ageing (TILDA). Report available on TILDA website: <https://tilda.tcd.ie/publications/reports/CohortMaintenance/> (Accessed February 2024).
- Dorner, T. E., J. Muckenhuber, W. J. Stronegger, E. Ràsky, B. Gustorff, and W. Freidl (2011). The impact of socio-economic status on pain and the perception of disability due to pain. *European Journal of Pain* 15(1), 103–109.
- Dreyer, L., S. Kendall, B. Danneskiold-Samsøe, E. M. Bartels, and H. Bliddal (2010). Mortality in a cohort of Danish patients with fibromyalgia: increased frequency of suicide. *Arthritis & Rheumatism* 62(10), 3101–3108.
- Dueñas, M., B. Ojeda, A. Salazar, J. A. Mico, and I. Failde (2016). A review of chronic pain impact on patients, their social environment and the health care system. *Journal of Pain Research* 9, 457–467.
- DuGoff, E. H., M. Schuler, and E. A. Stuart (2014). Generalizing observational study results: applying propensity score methods to complex surveys. *Health Services Research* 49(1), 284–303.
- Duncan, T. E. and S. C. Duncan (2009). The ABC's of LGM: An introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass* 3(6), 979–991.
- Duncan, T. E., S. C. Duncan, and L. A. Strycker (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Routledge.
- Dunn, K. M., K. Jordan, and P. R. Croft (2006). Characterizing the course of low back pain: a latent class analysis. *American Journal of Epidemiology* 163(8), 754–761.
- Dunn, K. M., K. P. Jordan, and P. R. Croft (2010). Recall of medication use, self-care activities and pain intensity: a comparison of daily diaries and self-report questionnaires among low back pain patients. *Primary Health Care Research & Development* 11(1), 93–102.
- Dziak, J. J., D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics* 21(2), 553–565.
- Dziak, J. J., S. T. Lanza, and X. Tan (2014). Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling: a Multidisciplinary Journal* 21(4), 534–552.

- D'Agostino McGowan, L. (2022). Sensitivity analyses for unmeasured confounders. *Current Epidemiology Reports* 9(4), 361–375.
- Ellenberg, J. H. (1994). Selection bias in observational and experimental studies. *Statistics in Medicine* 13(5-7), 557–567.
- Elo, I. T. (2009). Social class differentials in health and mortality: Patterns and explanations in comparative perspective. *Annual Review of Sociology* 35, 553–572.
- Emilsson, L., X. García-Albéniz, R. W. Logan, E. C. Caniglia, M. Kalager, and M. A. Hernán (2018). Examining bias in studies of statin treatment and survival in patients with cancer. *JAMA Oncology* 4(1), 63–70.
- Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science* 196(4286), 129–136.
- Fayaz, A., S. Ayis, S. S. Panesar, R. M. Langford, and L. J. Donaldson (2016). Assessing the relationship between chronic pain and cardiovascular disease: a systematic review and meta-analysis. *Scandinavian Journal of Pain* 13(1), 76–90.
- Fedak, K. M., A. Bernal, Z. A. Capshaw, and S. Gross (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, 1–9.
- Ferguson, K. D., M. McCann, S. V. Katikireddi, H. Thomson, M. J. Green, D. J. Smith, and J. D. Lewsey (2020). Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. *International Journal of Epidemiology* 49(1), 322–329.
- Fillingim, R. B., J. D. Loeser, R. Baron, and R. R. Edwards (2016). Assessment of chronic pain: domains, methods, and mechanisms. *The Journal of Pain* 17(9), T10–T20.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Number 5. Oliver and Boyd.
- Flegal, K. M., B. K. Kit, H. Orpana, and B. I. Graubard (2013). Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *Journal of the American Medical Association* 309(1), 71–82.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* 97(458), 611–631.

- Froehlich-Grobe, K., D. Jones, M. S. Businelle, D. E. Kendzor, and B. A. Balasubramanian (2016). Impact of disability and chronic conditions on health. *Disability and Health Journal* 9(4), 600–608.
- Fulcher, I. R., I. Shpitser, S. Marealle, and E. J. Tchetgen Tchetgen (2020). Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82(1), 199–214.
- Funk, M. J., D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology* 173(7), 761–767.
- Galea, S. and M. A. Hernán (2020). Win-win: reconciling social epidemiology and causal inference. *American Journal of Epidemiology* 189(3), 167–170.
- Gateway to Global Aging (2023). Gateway to Global Aging Data. <https://g2aging.org/index.php?section=concordance>. Accessed: 09 October 2023.
- Gellert, C., B. Schöttker, and H. Brenner (2012). Smoking and all-cause mortality in older people: systematic review and meta-analysis. *Archives of Internal Medicine* 172(11), 837–844.
- Gendreau, M., M. R. Hufford, and A. A. Stone (2003). Measuring clinical pain in chronic widespread pain: selected methodological issues. *Best Practice & Research Clinical Rheumatology* 17(4), 575–592.
- Gibson, S. J. (2007). IASP global year against pain in older persons: highlighting the current status and future perspectives in geriatric pain. *Expert Review of Neurotherapeutics* 7(6), 627–635.
- Glymour, M. M. and K. E. Rudolph (2016). Causal inference challenges in social epidemiology: bias, specificity, and imagination. *Social Science & Medicine* 100(166), 258–265.
- Gormley, I. C., T. B. Murphy, and A. E. Raftery (2023). Model-based clustering. *Annual Review of Statistics and Its Application* 10, 573–595.
- Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 14(3), 300–306.
- Greenland, S. and H. Morgenstern (2001). Confounding in health research. *Annual Review of Public Health* 22(1), 189–212.
- Greenland, S., J. Pearl, and J. M. Robins (1999). Confounding and collapsibility in causal inference. *Statistical Science* 14(1), 29–46.

BIBLIOGRAPHY

- Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31(4), 337–350.
- Greifer, N. (2023). cobalt: Covariate Balance Tables and Plots.
- Grol-Prokopczyk, H. (2017). Sociodemographic disparities in chronic pain, based on 12-year longitudinal data. *PAIN[®]* 158(2), 313–322.
- Grol-Prokopczyk, H., E. Verdes-Tennant, M. McEniry, and M. Ispány (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography* 52(5), 1703–1728.
- Grün, B. (2019). Model-based clustering. In *Handbook of Mixture Analysis*, pp. 157–192. Chapman and Hall/CRC.
- Haber, N. A., S. E. Wieten, J. M. Rohrer, O. A. Arah, P. W. Tennant, E. A. Stuart, E. J. Murray, S. Pilleron, S. T. Lam, and E. Riederer (2022). Causal and associational language in observational health research: a systematic evaluation. *American Journal of Epidemiology* 191(12), 2084–2097.
- Hadi, M. A., G. A. McHugh, and S. J. Closs (2019). Impact of chronic pain on patients’ quality of life: a comparative mixed-methods study. *Journal of Patient Experience* 6(2), 133–141.
- Hammerton, G. and M. R. Munafò (2021). Causal inference with observational data: the need for triangulation of evidence. *Psychological Medicine* 51(4), 563–578.
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 99(467), 609–618.
- Hartge, P. (2006). Participation in population studies. *Epidemiology* 17(3), 252–254.
- Health and Retirement Study (2023). 1998-2018 public use datasets. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740).
- Health Insurance Authority (2023). Health Insurance in Ireland Market Report 2021. Report available online: https://www.hia.ie/sites/default/files/2023-01/hia-market-report-2021_1.pdf (Accessed May 2023).
- Hernán, M. A. (2012). Beyond exchangeability: the other conditions for causal inference in medical research. *Statistical Methods in Medical Research* 21(1), 3–5.

- Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology* 26(10), 674–680.
- Hernán, M. A. (2018). The C-word: scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health* 108(5), 616–619.
- Hernán, M. A. and S. R. Cole (2009). Invited commentary: causal diagrams and measurement bias. *American Journal of Epidemiology* 170(8), 959–962.
- Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004). A structural approach to selection bias. *Epidemiology* 15(5), 615–625.
- Hernán, M. A., S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* 155(2), 176–184.
- Hernán, M. A. and J. M. Robins (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health* 60(7), 578–586.
- Hernán, M. A. and J. M. Robins (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology* 183(8), 758–764.
- Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Hernández-Díaz, S., E. F. Schisterman, and M. A. Hernán (2006). The birth weight “paradox” uncovered? *American Journal of Epidemiology* 164(11), 1115–1120.
- Hicks, G. E., J. M. Sions, P. C. Coyle, and R. T. Pohlig (2017). Altered spatiotemporal characteristics of gait in older adults with chronic low back pain. *Gait & Posture* 55, 172–176.
- Holt-Lunstad, J., T. B. Smith, M. Baker, T. Harris, and D. Stephenson (2015). Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on Psychological Science* 10(2), 227–237.
- Hooper, D., J. Coughlan, and M. Mullen (2008). Evaluating model fit: a synthesis of the structural equation modelling literature. In *7th European Conference on Research Methodology for Business and Management Studies*, pp. 195–200.
- Hooten, W. M. (2016). Chronic pain and mental health disorders: shared neural mechanisms, epidemiology, and treatment. *Mayo Clinic Proceedings* 91(7), 955–970.

- Howe, L. D., K. Tilling, B. Galobardes, and D. A. Lawlor (2013). Loss to follow-up in cohort studies: Bias in estimates of socioeconomic inequalities. *Epidemiology* 24(1), 1–9.
- Hu, L.-t. and P. M. Bentler (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 6(1), 1–55.
- Idler, E., J. Blevins, M. Kiser, and C. Hogue (2017). Religion, a social determinant of mortality? a 10-year follow-up of the health and retirement study. *PloS One* 12(12), e0189134.
- Ikeda, T., K. Sugiyama, J. Aida, T. Tsuboya, N. Watabiki, K. Kondo, and K. Osaka (2019). Socioeconomic inequalities in low back pain among older people: the JAGES cross-sectional study. *International Journal for Equity in Health* 18(1), 1–11.
- Inoue, K., B. Ritz, and O. A. Arah (2022). Causal effect of chronic pain on mortality through opioid prescriptions: Application of the front-door formula. *Epidemiology (Cambridge, Mass.)* 33(4), 572.
- Institute for Social Research University of Michigan (2023). The Health and Retirement Study Documentation. <https://hrs.isr.umich.edu/documentation>. (Accessed: January 2024).
- Jacobs, J. M., R. Hammerman-Rozenberg, A. Cohen, and J. Stessman (2006). Chronic back pain among the elderly: prevalence, associations, and predictors. *Spine* 31(7), E203–E207.
- James, G., D. Witten, T. Hastie, R. Tibshirani, et al. (2013). *An introduction to statistical learning*, Volume 112. Springer.
- Jurek, A. M., G. Maldonado, S. Greenland, and T. R. Church (2006). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European Journal of Epidemiology* 21, 871–876.
- Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics* 16(2), 163–178.
- Kadane, J. B. and N. A. Lazar (2004). Methods and criteria for model selection. *Journal of the American Statistical Association* 99(465), 279–290.
- Karim, J., R. Weisz, Z. Bibi, and S. ur Rehman (2015). Validation of the eight-item center for epidemiologic studies depression scale (CES-D) among older adults. *Current Psychology* 34, 681–692.
- Kassambara, A., M. Kosinski, and P. Biecek (2021). survminer: Drawing survival curves using 'ggplot2'. R package version 0.4.9. <https://CRAN.R-project.org/package=survminer>.

- Kawai, K., A. T. Kawai, P. Wollan, and B. P. Yawn (2017). Adverse impacts of chronic pain on health-related quality of life, work productivity, depression and anxiety in a community-based study. *Family Practice* 34(6), 656–661.
- Kendall, J. (2003). Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal* 20(2), 164–168.
- Kennedy, N., K. O’Sullivan, A. Hannigan, and H. Purtill (2017). Understanding pain among older persons: Part 2—the association between pain profiles and healthcare utilisation. *Age and Ageing* 46(1), 51–56.
- Kenny, R. A., B. J. Whelan, H. Cronin, Y. Kamiya, P. Kearney, C. O’Regan, and M. Ziegel (2010). The design of the Irish Longitudinal Study on Ageing. Technical report, The Irish Longitudinal Study on Ageing (TILDA). https://tilda.tcd.ie/publications/reports/pdf/Report_DesignReport.pdf.
- Kleinbaum, D. G. and M. Klein (1996). *Survival analysis a self-learning text*. Springer.
- Kongsted, A., P. Kent, I. Axen, A. S. Downie, and K. M. Dunn (2016). What have we learned from ten years of trajectory research in low back pain? *BMC Musculoskeletal Disorders* 17, 1–11.
- Kreif, N., R. Grieve, R. Radice, and J. S. Sekhon (2013). Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services and Outcomes Research Methodology* 13, 174–202.
- Krieger, N. and G. Davey Smith (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.
- Kroenke, K., J. Wu, M. J. Bair, E. E. Krebs, T. M. Damush, and W. Tu (2011). Reciprocal relationship between pain and depression: a 12-month longitudinal analysis in primary care. *The Journal of Pain* 12(9), 964–973.
- Kvale, E., O. J. Ekundayo, Y. Zhang, S. Akhter, I. Aban, T. E. Love, C. Ritchie, and A. Ahmed (2011). History of cancer and mortality in community-dwelling older adults. *Cancer Epidemiology* 35(1), 30–36.
- Lacey, R. J., J. Belcher, and P. R. Croft (2013). Does life course socio-economic position influence chronic disabling pain in older adults? A general population study. *The European Journal of Public Health* 23(4), 534–540.
- Lacey, R. J., K. P. Jordan, and P. R. Croft (2013). Does attrition during follow-up of a population cohort study inevitably lead to biased estimates of health status? *PLoS One* 8(12), e83948.

- Landmark, T., P. Romundstad, O. Dale, P. C. Borchgrevink, and S. Kaasa (2012). Estimating the prevalence of chronic pain: Validation of recall against longitudinal reporting (the HUNT pain study). *PAIN[®]* 153(7), 1368–1373.
- Lanza, S. T., D. L. Coffman, and S. Xu (2013). Causal inference in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal* 20(3), 361–383.
- Lanza, S. T., L. M. Collins, D. R. Lemmon, and J. L. Schafer (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling: a Multidisciplinary Journal* 14(4), 671–694.
- Lanza, S. T., B. P. Flaherty, and L. M. Collins (2003). Latent class and latent transition analysis. In *Handbook of Psychology*, pp. 663–685. John Wiley & Sons.
- Lazarsfeld, P. and N. Henry (1968). *Latent Structure Analysis*. Houghton, Mifflin.
- Leadley, R. M., N. Armstrong, K. J. Reid, A. Allen, K. V. Misso, and J. Kleijnen (2014). Healthy aging in relation to chronic pain and quality of life in Europe. *Pain Practice* 14(6), 547–558.
- Liang, Y. (2024). Association between chronic pain and attrition: a prospective analysis of a national sample of midlife adults in the USA, 2004–2014. *BMJ Public Health* 2(1), e000564.
- Lima-Costa, M. F., J. V. de Melo Mambrini, F. Bof de Andrade, P. R. B. de Souza Jr, M. T. L. De Vasconcellos, A. L. Neri, E. Castro-Costa, J. Macinko, and C. De Oliveira (2023). Cohort profile: The Brazilian Longitudinal Study of Ageing (ELSI-Brazil). *International Journal of Epidemiology* 52(1), e57–e65.
- Lin, M., H. C. Lucas Jr, and G. Shmueli (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research* 24(4), 906–917.
- Linzer, D. A. and J. B. Lewis (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software* 42, 1–29.
- Lu, H., S. R. Cole, C. J. Howe, and D. Westreich (2022). Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology* 33(5), 699–706.
- Lubke, G. H. and B. Muthén (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods* 10(1), 21.

- Ludwig, C., C. Luthy, A. Allaz, F. Herrmann, and C. Cedraschi (2018). The impact of low back pain on health-related quality of life in old age: results from a survey of a large sample of Swiss elders living in the community. *European Spine Journal* 27, 1157–1165.
- Lythgoe, D. T., M. Garcia-Fiñana, and T. F. Cox (2019). Latent class modeling with a time-to-event distal outcome: A comparison of one, two and three-step approaches. *Structural Equation Modeling: A Multidisciplinary Journal* 26(1), 51–65.
- Ma, J., E. M. Ward, R. L. Siegel, and A. Jemal (2015). Temporal trends in mortality in the United States, 1969–2013. *Journal of the American Medical Association* 314(16), 1731–1739.
- Macfarlane, G. J., G. Jones, P. Knekt, A. Aromaa, J. McBeth, M. Mikkelsen, and M. Heliovaara (2007). Is the report of widespread body pain associated with long-term increased mortality? Data from the Mini-Finland Health Survey. *Rheumatology* 46(5), 805–807.
- Macfarlane, G. J., J. McBeth, and A. J. Silman (2001). Widespread body pain and mortality: prospective population based study. *British Medical Journal* 323, 1–5.
- Makino, K., S. Lee, S. Lee, S. Bae, S. Jung, Y. Shinkai, and H. Shimada (2019). Daily physical activity and functional disability incidence in community-dwelling older adults with chronic pain: a prospective cohort study. *Pain Medicine* 20(9), 1702–1710.
- Makris, U. E., L. Fraenkel, L. Han, L. Leo-Summers, and T. M. Gill (2014). Restricting back pain and subsequent mobility disability in community-living older persons. *Journal of the American Geriatrics Society* 62(11), 2142–2147.
- Makris, U. E., R. T. Higashi, E. G. Marks, L. Fraenkel, T. M. Gill, J. L. Friedly, and M. C. Reid (2017). Physical, emotional, and social impacts of restricting back pain in older adults: A qualitative study. *Pain Medicine* 18(7), 1225–1235.
- Mangiafico, S. (2022). rcompanion: Functions to Support Extension Education Program Evaluation. R package version 2.4.18. <https://CRAN.R-project.org/package=rcompanion>.
- Mansournia, M. A., M. A. Hernán, and S. Greenland (2013). Matched designs and causal diagrams. *International Journal of Epidemiology* 42(3), 860–869.
- Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*, pp. 551–611. Oxford University Press.

- Masyn, K. E., H. Petras, and W. Liu (2014). Growth curve models with categorical outcomes. In *Encyclopedia of Criminology and Criminal Justice*, pp. 2013–2025. Wiley-Blackwell.
- McBeth, J., D. Symmons, A. Silman, T. Allison, R. Webb, T. Brammah, and G. Macfarlane (2009). Musculoskeletal pain is associated with a long-term increased risk of cancer and cardiovascular-related mortality. *Rheumatology* 48(1), 74–77.
- McFadden, D. (1979). Quantitative methods for analysing travel behaviour of individuals: some recent developments. In *Behavioural Travel Modelling*, pp. 279–318. Routledge.
- McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley.
- Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal* 2, 769–782.
- Mehta, P. D., M. C. Neale, and B. R. Flay (2004). Squeezing interval change from ordinal panel data: latent growth curves with ordinal outcomes. *Psychological Methods* 9(3), 301.
- Melchiorre, M. G., C. Chiatti, G. Lamura, F. Torres-Gonzales, M. Stankunas, J. Lindert, E. Ioannidi-Kapolou, H. Barros, G. Macassa, and J. F. Soares (2013). Social support, socio-economic status, health and abuse among older people in seven European countries. *PloS One* 8(1), e54856.
- Menec, V. H., S. Shooshtari, and P. Lambert (2007). Ethnic differences in self-rated health among older adults: A cross-sectional and longitudinal analysis. *Journal of Aging and Health* 19(1), 62–86.
- Metten, M.-A., N. Costet, L. Multigner, J.-F. Viel, and G. Chauvet (2022). Inverse probability weighting to handle attrition in cohort studies: some guidance and a call for caution. *BMC Medical Research Methodology* 22(1), 1–15.
- Miaskowski, C., F. Blyth, F. Nicosia, M. Haan, F. Keefe, A. Smith, and C. Ritchie (2020). A biopsychosocial model of chronic pain for older adults. *Pain Medicine* 21(9), 1793–1805.
- Milani, S. A., B. Howrey, M. A. Rodriguez, R. Samper-Ternent, and R. Wong (2022). Gender differences in activity-limiting pain trajectories over a 17-year period in the Mexican Health and Aging Study. *PAIN[®]* 163(2), e285.
- Mills, S. E., K. P. Nicolson, and B. H. Smith (2019). Chronic pain: a review of its epidemiology and associated factors in population-based studies. *British Journal of Anaesthesia* 123(2), e273–e283.

- Mohanty, S. K., M. Ambade, A. K. Upadhyay, R. S. Mishra, S. P. Pedgaonkar, F. Kampfen, O. O'Donnell, and J. Maurer (2022). Prevalence of pain and its treatment among older adults in India: a nationally representative population-based study. *PAIN[®]* 164(2), 336–348.
- Molina, T. (2016). Reporting heterogeneity and health disparities across gender and education levels: Evidence from four countries. *Demography* 53(2), 295–323.
- Morales, M. E. and R. J. Yong (2021). Racial and ethnic disparities in the treatment of chronic pain. *Pain Medicine* 22(1), 75–90.
- Mortimer, K. M., R. Neugebauer, M. Van Der Laan, and I. B. Tager (2005). An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology* 162(4), 382–388.
- Mullins, P. M., R. J. Yong, and N. Bhattacharyya (2022). Impact of demographic factors on chronic pain among adults in the United States. *Pain Reports* 7(4), e1009.
- Musich, S., S. S. Wang, J. Ruiz, K. Hawkins, and E. Wicker (2018). The impact of mobility limitations on health outcomes among older adults. *Geriatric Nursing* 39(2), 162–169.
- Muszyńska-Spielauer, M. and M. Spielauer (2022). Cross-sectional estimates of population health from the Survey of Health and Retirement in Europe (SHARE) are biased due to health-related sample attrition. *SSM-Population Health* 20, 101290.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika* 29, 81–117.
- Naimi, A. I., S. R. Cole, and E. H. Kennedy (2017). An introduction to g methods. *International Journal of Epidemiology* 46(2), 756–762.
- Neogi, T. (2013). The epidemiology and impact of pain in osteoarthritis. *Osteoarthritis and Cartilage* 21(9), 1145–1153.
- Nitter, A. K. and K. Ø. Forseth (2013). Mortality rate and causes of death in women with self-reported musculoskeletal pain: Results from a 17-year follow-up study. *Scandinavian Journal of Pain* 4(2), 86–92.
- Nohr, E. A. and Z. Liew (2018). How to investigate and adjust for selection bias in cohort studies. *Acta Obstetrica et Gynecologica Scandinavica* 97(4), 407–416.
- Nylund, K. L., T. Asparouhov, and B. O. Muthén (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal* 14(4), 535–569.

- Nylund-Gibson, K. and A. Y. Choi (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science* 4(4), 440.
- Nylund-Gibson, K., A. C. Garber, D. B. Carter, M. Chan, D. A. Arch, O. Simon, K. Whaling, E. Tartt, and S. I. Lawrie (2023). Ten frequently asked questions about latent transition analysis. *Psychological Methods* 28(2), 284.
- Oberski, D. (2023). poLCA.extras: Some additional functionality for poLCA models. R package version 0.1.0.
- Ojala, T., A. Häkkinen, J. Karppinen, K. Sipilä, T. Suutama, and A. Piirainen (2015). Chronic pain affects the whole person—a phenomenological study. *Disability and Rehabilitation* 37(4), 363–371.
- Olshansky, S. J., T. Antonucci, L. Berkman, R. H. Binstock, A. Boersch-Supan, J. T. Cacioppo, B. A. Carnes, L. L. Carstensen, L. P. Fried, D. P. Goldman, et al. (2012). Differences in life expectancy due to race and educational differences are widening, and many may not catch up. *Health Affairs* 31(8), 1803–1813.
- O’Neill, A., K. O’Sullivan, M. O’Keeffe, A. Hannigan, C. Walsh, and H. Purtill (2018). Development of pain in older adults: A latent class analysis of biopsychosocial risk factors. *PAIN[®]* 159(8), 1631–1640.
- O’Neill, A., K. O’Sullivan, M. O’Keeffe, C. Walsh, and H. Purtill (2020). The change of pain classes over time: a latent transition analysis. *European Journal of Pain* 24(2), 457–469.
- Palacios-Ceña, D., C. Alonso-Blanco, V. Hernández-Barrera, P. Carrasco-Garrido, R. Jiménez-García, and C. Fernández-de-las Peñas (2015). Prevalence of neck and low back pain in community-dwelling adults in Spain: an updated population-based national study (2009/10–2011/12). *European Spine Journal* 24, 482–492.
- Parsons, S., J. McBeth, G. Macfarlane, P. Hannaford, and D. Symmons (2015). Self-reported pain severity is associated with a history of coronary heart disease. *European Journal of Pain* 19(2), 167–175.
- Pearl, J. (1993). Comment: graphical models, causality and intervention. *Statistical Science* 8(3), 266–269.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* 82(4), 669–688.
- Pearl, J. (2010a). Brief report: On the consistency rule in causal inference: “Axiom, definition, assumption, or theorem?”. *Epidemiology* 21(6), 872–875.

- Pearl, J. (2010b). The foundations of causal inference. *Sociological Methodology* 40(1), 75–149.
- Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. Van Der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21(1), 31–54.
- Pillay, T., H. A. Van Zyl, and D. Blackbeard (2014). Chronic pain perception and cultural experience. *Procedia-Social and Behavioral Sciences* 113, 151–160.
- Pishgar, F., N. Greifer, C. Leyrat, and E. Stuart (2021). MatchThem:: Matching and Weighting after Multiple Imputation. *The R Journal* 13(2), 292–305.
- Pohle, J., R. Langrock, F. M. Van Beest, and N. M. Schmidt (2017). Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics* 22, 270–293.
- Preacher, K. J., A. L. Wichman, R. C. MacCallum, and N. E. Briggs (2008). *Latent Growth Curve Modeling*. Sage.
- Qian, M., Y. Shi, and M. Yu (2021). The association between obesity and chronic pain among community-dwelling older adults: a systematic review and meta-analysis. *Geriatric Nursing* 42(1), 8–15.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raftery, A. E. (1995). Bayesian model selection in social research. In *Sociological Methodology*, pp. 111–163. American Sociological Association.
- Raftery, M. N., K. Sarma, A. W. Murphy, D. De la Harpe, C. Normand, and B. E. McGuire (2011). Chronic pain in the Republic of Ireland—community prevalence, psychosocial profile and predictors of pain-related disability: results from the Prevalence, Impact and Cost of Chronic Pain (PRIME) study, part 1. *PAIN[®]* 152(5), 1096–1103.
- Raja, S. N., D. B. Carr, M. Cohen, N. B. Finnerup, H. Flor, S. Gibson, F. J. Keefe, J. S. Mogil, M. Ringkamp, K. A. Sluka, et al. (2020). The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *PAIN[®]* 161(9), 1976–1982.
- Ray, L., R. B. Lipton, M. E. Zimmerman, M. J. Katz, and C. A. Derby (2011). Mechanisms of association between obesity and chronic pain in the elderly. *PAIN[®]* 152(1), 53–59.

- Ray, W. A., C. P. Chung, K. T. Murray, K. Hall, and C. M. Stein (2016). Prescription of long-acting opioids and mortality in patients with chronic noncancer pain. *Journal of the American Medical Association* 315(22), 2415–2423.
- Rehkopf, D. H., M. M. Glymour, and T. L. Osypuk (2016). The consistency assumption for causal inference in social epidemiology: when a rose is not a rose. *Current Epidemiology Reports* 3(1), 63–71.
- Rice, A. S., B. H. Smith, and F. M. Blyth (2016). Pain and the global burden of disease. *PAIN[®]* 157(4), 791–796.
- Ridgeway, G., S. A. Kovalchik, B. A. Griffin, and M. U. Kabeto (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference* 3(2), 237–249.
- Robins, J. and M. Hernán (2008). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, pp. 553–599. Chapman and Hall/CRC.
- Robins, J. M. and M. B. Weissman (2016). Commentary: Counterfactual causation and streetlamps: what is to be done? *International Journal of Epidemiology* 45(6), 1830–1835.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48, 1–36.
- Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* 93(444), 1321–1339.
- Rubin, D. (1987). *Multiple Imputation for Non-response in Surveys*, Volume 10. New York: John Wiley & Sons, Inc.
- Ryan, E., J. J. Dziak, H. Purtill, and B. C. Bray (2023). Can a Normed Fit Index Assist with Model Selection in Latent Class Analysis with Large Samples? A Preliminary Investigation. Technical report available online: <https://doi.org/10.31234/osf.io/3qzvm>.
- Ryan, E., H. Grol-Prokopczyk, C. R. Dennison, A. Zajacova, and Z. Zimmer (2024). Is the relationship between chronic pain and mortality causal? A propensity score analysis. *PAIN[®]*. <https://doi.org/10.1097/j.pain.0000000000003336>.
- Ryan, E., A. Hannigan, H. Grol-Prokopczyk, P. May, and H. Purtill (2024). Sociodemographic disparities and potential biases in persistent pain estimates: Findings from 5 waves of the Irish Longitudinal Study on Ageing (TILDA). *European Journal of Pain* 28(5), 754–768.

- Sánchez, B. N., E. Budtz-Jørgensen, L. M. Ryan, and H. Hu (2005). Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association* 100(472), 1443–1455.
- Schisterman, E. F., S. R. Cole, and R. W. Platt (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 20(4), 488–495.
- Schoevers, R., M. Geerlings, D. Deeg, T. Holwerda, C. Jonker, and A. Beekman (2009). Depression and excess mortality: evidence for a dose response relation in community living elderly. *International Journal of Geriatric Psychiatry* 24(2), 169–176.
- Schratz, P. (2017). R package 'oddsratio': Odds ratio calculation for GAM(M)s & GLM(M)s. R package version 1.0.2. doi: 10.5281/zenodo.1095472.
- Schuler, M. S., J.-M. S. Leoutsakos, and E. A. Stuart (2014). Addressing confounding when estimating the effects of latent classes on a distal outcome. *Health Services and Outcomes Research Methodology* 14, 232–254.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343.
- Sheng, J., S. Liu, Y. Wang, R. Cui, and X. Zhang (2017). The link between depression and chronic pain: neural mechanisms in the brain. *Neural Plasticity* 2017(1), 9724371.
- Sher, K. J., K. M. Jackson, and D. Steinley (2011). Alcohol use trajectories and the ubiquitous cat's cradle: cause for concern? *Journal of Abnormal Psychology* 120(2), 322.
- Shimonovich, M., A. Pearce, H. Thomson, K. Keyes, and S. V. Katikireddi (2021). Assessing causality in epidemiology: revisiting Bradford Hill to incorporate developments in causal thinking. *European Journal of Epidemiology* 36, 873–887.
- Shupler, M. S., J. K. Kramer, J. J. Cragg, C. R. Jutzeler, and D. G. Whitehurst (2019). Pan-Canadian estimates of chronic pain prevalence from 2000 to 2014: a repeated cross-sectional survey analysis. *The Journal of Pain* 20(5), 557–565.
- Simmonds, M. J., C. E. Lee, B. R. Etnyre, and G. S. Morris (2012). The influence of pain distribution on walking velocity and horizontal ground reaction forces in patients with low back pain. *Pain Research and Treatment* 2012, 1–10.

- Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11), 1359–1366.
- Sjøgren, P., O. Ekholm, V. Peuckmann, and M. Grønbaek (2009). Epidemiology of chronic pain in Denmark: an update. *European Journal of Pain* 13(3), 287–292.
- Smith, A. and K. Dunn (2022). Research Note: Deriving latent trajectories in health research. *Journal of Physiotherapy* 69, 61–64.
- Smith, B. H., A. M. Elliott, and P. C. Hannaford (2003). Pain and subsequent mortality and cancer among women in the Royal College of General Practitioners Oral Contraception Study. *British Journal of General Practice* 53(486), 45–46.
- Smith, D., R. Wilkie, P. Croft, and J. McBeth (2018). Pain and mortality in older adults: the influence of pain phenotype. *Arthritis Care & Research* 70(2), 236–243.
- Smith, D., R. Wilkie, P. Croft, S. Parmar, and J. McBeth (2018). Pain and mortality: mechanisms for a relationship. *PAIN[®]* 159(6), 1112–1118.
- Smith, D., R. Wilkie, O. Uthman, J. L. Jordan, and J. McBeth (2014). Chronic pain and mortality: a systematic review. *PloS One* 9(6), e99048.
- Smith, L. H. (2020). Selection mechanisms and their consequences: Understanding and addressing selection bias. *Current Epidemiology Reports* 7, 179–189.
- Sonnega, A., J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. Phillips, and D. R. Weir (2014). Cohort profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology* 43(2), 576–585.
- Spencer, J. C., S. B. Wheeler, J. S. Rotter, and G. M. Holmes (2018). Decomposing mortality disparities in urban and rural US counties. *Health Services Research* 53(6), 4310–4331.
- Spitzer, S. and D. Weber (2019). Reporting biases in self-assessed physical and cognitive health status of older Europeans. *PLoS One* 14(10), e0223526.
- Stephens, A., E. Breeze, J. Banks, and J. Nazroo (2013). Cohort profile: the English Longitudinal Study of Ageing. *International Journal of Epidemiology* 42(6), 1640–1648.
- Stewart Williams, J., N. Ng, K. Peltzer, A. Yawson, R. Biritwum, T. Maximova, F. Wu, P. Arokiasamy, P. Kowal, and S. Chatterji (2015). Risk factors and disability associated with low back pain in older adults in low-and middle-income countries. Results from the WHO Study on Global AGEing and Adult Health (SAGE). *PloS One* 10(6), e0127880.

- Stynes, S., K. Konstantinou, R. Ogollah, E. M. Hay, and K. M. Dunn (2018). Novel approach to characterising individuals with low back-related leg pain: cluster identification with latent class analysis and 12-month follow-up. *PAIN[®]* 159(4), 728–738.
- Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* 12(3), e1001779.
- Sullivan, M. D. and J. C. Ballantyne (2016). Must we reduce pain intensity to treat chronic pain? *PAIN[®]* 157(1), 65–69.
- Surah, A., G. Baranidharan, and S. Morley (2014). Chronic pain and depression. *Continuing Education in Anaesthesia, Critical Care & Pain* 14(2), 85–89.
- Sutradhar, R. and P. C. Austin (2018). Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology* 28(1), 54–57.
- Szklo, M. (1998). Population-based cohort studies. *Epidemiologic Reviews* 20(1), 81–90.
- Tchetgen Tchetgen, E. J. and K. E. Wirth (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* 73(4), 1123–1131.
- Tennant, P. W., E. J. Murray, K. F. Arnold, L. Berrie, M. P. Fox, S. C. Gadd, W. J. Harrison, C. Keeble, L. R. Ranker, and J. Textor (2021). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology* 50(2), 620–632.
- The Irish Longitudinal Study on Ageing (2024). Waves 1-5, 2009-2018. [dataset]. Irish Social Science Data Archive.
- Therneau, T. M. (2022). A Package for Survival Analysis in R. R package version 3.4-0. <https://CRAN.R-project.org/package=survival>.
- Thienhaus, O. and B. E. Cole (2002). Classification of pain. In *Pain management: A practical guide for clinicians*, pp. 27–36. CRC press Boca Raton, FL.
- Tobias, D. K. and F. B. Hu (2018). The association between BMI and mortality: implications for obesity prevention. *The Lancet Diabetes & Endocrinology* 6(12), 916–917.

- Torrance, N., A. M. Elliott, A. J. Lee, and B. H. Smith (2010). Severe chronic pain is associated with increased 10 year mortality. A cohort record linkage study. *European Journal of Pain* 14(4), 380–386.
- Tucker, L. R. and C. Lewis (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38(1), 1–10.
- Turner, B., S. Smith, and S. Thomson (2020). Uncovering the complex role of private health insurance in Ireland. In *Private Health Insurance: History, Politics and Performance*, pp. 221–63. Cambridge University Press.
- Tölle, T., M. Fitzcharles, and W. Häuser (2021). Is opioid therapy for chronic non-cancer pain associated with a greater risk of all-cause mortality compared to non-opioid analgesics? a systematic review of propensity score matched observational studies. *European Journal of Pain* 25(6), 1195–1208.
- Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (Second Edition ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45(3), 1–67.
- Van Eenoo, L., A. Declercq, G. Onder, H. Finne-Soveri, V. Garms-Homolova, P. V. Jonsson, O. H. Dix, J. H. Smit, H. P. Van Hout, and H. G. Van Der Roest (2016). Substantial between-country differences in organising community care for older people in Europe—a review. *The European Journal of Public Health* 26(2), 213–219.
- Van Hecke, O., N. Torrance, and B. Smith (2013). Chronic pain epidemiology and its clinical relevance. *British Journal of Anaesthesia* 111(1), 13–18.
- Van Smeden, M., T. L. Lash, and R. H. Groenwold (2020). Reflection on modern methods: five myths about measurement error in epidemiological research. *International Journal of Epidemiology* 49(1), 338–347.
- Van Zwieten, A., P. W. Tennant, M. Kelly-Irving, F. M. Blyth, A. Teixeira-Pinto, and S. Khalatbari-Soltani (2022). Avoiding overadjustment bias in social epidemiology through appropriate covariate selection: a primer. *Journal of Clinical Epidemiology* 149, 127–136.
- Vandenbroucke, J. P., A. Broadbent, and N. Pearce (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology* 45(6), 1776–1786.
- VanderWeele, T. and S. Vansteelandt (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods* 2(1), 95–115.

- VanderWeele, T. J. (2009a). Concerning the consistency assumption in causal inference. *Epidemiology* 20(6), 880–883.
- VanderWeele, T. J. (2009b). On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology* 20(4), 496–499.
- VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology* 34, 211–219.
- VanderWeele, T. J. (2022). Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs. *Epidemiology (Cambridge, Mass.)* 33(1), 141.
- VanderWeele, T. J. and P. Ding (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine* 167(4), 268–274.
- VanderWeele, T. J. and M. A. Hernán (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference* 1(1), 1–20.
- Vansteelandt, S. and R. M. Daniel (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology* 28(2), 258–265.
- Vermunt, J. K. and J. Magidson (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. Technical report, Statistical Innovations Inc, Belmont, MA.
- Vermunt, J. K. and J. Magidson (2021). LG-Syntax User’s Guide: Manual for Latent Gold Syntax Module Version 6.0. Technical report, Statistical Innovations Inc, Arlington, MA.
- Visser, M. and S. Depaoli (2022). A guide to detecting and modeling local dependence in latent class analysis models. *Structural Equation Modeling: A Multidisciplinary Journal* 29(6), 971–982.
- Wade, K. F., A. Marshall, B. Vanhoutte, F. C. Wu, T. W. O’Neill, and D. M. Lee (2017). Does pain predict frailty in older men and women? Findings from the English Longitudinal Study of Ageing (ELSA). *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 72(3), 403–409.
- Wade, K. H., D. Carslake, N. Sattar, G. Davey Smith, and N. J. Timpson (2018). BMI and mortality in UK Biobank: revised estimates using Mendelian randomization. *Obesity* 26(11), 1796–1806.
- Walsh, D. A. and D. F. McWilliams (2014). Mechanisms, impact and management of pain in rheumatoid arthritis. *Nature Reviews Rheumatology* 10(10), 581–592.

- Walters, G. D. (2011). The latent structure of life-course-persistent antisocial behavior: Is Moffitt's developmental taxonomy a true taxonomy? *Journal of Consulting and Clinical Psychology* 79(1), 96.
- Wang, H., L. Dwyer-Lindgren, K. T. Lofgren, J. K. Rajaratnam, J. R. Marcus, A. Levin-Rector, C. E. Levitz, A. D. Lopez, and C. J. Murray (2012). Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 380(9859), 2071–2094.
- Ward, M., P. May, R. Briggs, T. McNicholas, C. Normand, R. A. Kenny, and A. Nolan (2020). Linking death registration and survey data: procedures and cohort profile for the Irish Longitudinal Study on Ageing. *HRB Open Research* 3, 1–54.
- Wasserstein, R. L. and N. A. Lazar (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician* 70(2), 129–133.
- Weir, D. R. (2016). Validating mortality ascertainment in the Health and Retirement Study. Technical report, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.
- Wendt, L. P., A. G. Wright, P. A. Pilkonis, T. Nolte, P. Fonagy, P. R. Montague, C. Benecke, T. Krieger, and J. Zimmermann (2019). The latent structure of interpersonal problems: Validity of dimensional, categorical, and hybrid models. *Journal of Abnormal Psychology* 128(8), 823.
- Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology (Cambridge, Mass.)* 23(1), 159.
- Westreich, D. and S. R. Cole (2010). Invited commentary: positivity in practice. *American Journal of Epidemiology* 171(6), 674–677.
- White, J., P. Zaninotto, K. Walters, M. Kivimäki, P. Demakakos, J. Biddulph, M. Kumari, C. De Oliveira, J. Gallacher, and G. D. Batty (2016). Duration of depressive symptoms and mortality risk: the English Longitudinal Study of Ageing (ELSA). *The British Journal of Psychiatry* 208(4), 337–342.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wickham, H., M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* 4(43), 1686.

- Wolfe, F., A. L. Hassett, B. Walitt, and K. Michaud (2011). Mortality in fibromyalgia: A study of 8,186 patients over thirty-five years. *Arthritis Care & Research* 63(1), 94–101.
- Wong, C. K., R. Y. Mak, T. S. Kwok, J. S. Tsang, M. Y. Leung, M. Funabashi, L. G. Macedo, L. Dennett, and A. Y. Wong (2022). Prevalence, incidence, and factors associated with non-specific chronic low back pain in community-dwelling older adults aged 60 years and older: a systematic review and meta-analysis. *The Journal of Pain* 23(4), 509–534.
- Wrangler, L. S., M. Rennemark, and J. Berglund (2016). Pain among older adults from a gender perspective: findings from the Swedish National Study on Aging and Care (SNAC-Blekinge). *Scandinavian Journal of Public Health* 44(3), 258–263.
- Yamaguchi, K. (2015). Extensions of Rubin’s Causal Model for a Latent-Class Treatment Variable: An analysis of the effects of employers’ work-life balance policies on women’s income attainment in Japan. Technical report, Research Institute of Economy, Trade and Industry (RIETI).
- Yoshida, K. and A. Bartel (2022). tableone: Create ‘Table 1’ to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.13.2.
- Zajacova, A., H. Grol-Prokopczyk, M. Limani, C. Schwarz, and I. Gilron (2023). Prevalence and correlates of prescription opioid use among US adults, 2019–2020. *PLoS One* 18(3), e0282536.
- Zajacova, A., H. Grol-Prokopczyk, and Z. Zimmer (2021). Pain trends among American adults, 2002–2018: patterns, disparities, and correlates. *Demography* 58(2), 711–738.
- Zajacova, A., J. Lee, and H. Grol-Prokopczyk (2022). The geography of pain in the United States and Canada. *The Journal of Pain* 23(12), 2155–2166.
- Zhao, Y., Y. Hu, J. P. Smith, J. Strauss, and G. Yang (2014). Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). *International Journal of Epidemiology* 43(1), 61–68.
- Zhu, Y., R. A. Hubbard, J. Chubak, J. Roy, and N. Mitra (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and Drug Safety* 30(11), 1471–1485.
- Ziebarth, N. (2010). Measurement of health, health inequality, and reporting heterogeneity. *Social Science & Medicine* 71(1), 116–124.

BIBLIOGRAPHY

Zimmer, Z., A. Zajacova, and H. Grol-Prokopczyk (2020). Trends in pain prevalence among adults aged 50 and older across Europe, 2004 to 2015. *Journal of Aging and Health* 32(10), 1419–1432.