Health perspectives: Exploring differential reporting across sex and generations

Alejandra Arrieta

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Emmanuela Gakidou, Chair

Luisa Flor

Heidi Stöckl

Theo Vos

Program Authorized to Offer Degree:

Health Metrics Sciences

University of Washington

Abstract

Health perspectives: Exploring differential reporting across sex and generations

Alejandra Arrieta

Chair of Supervisory Committee:

Emmanuela Gakidou

Department of Health Metrics Sciences

In health metrics, health surveys are an important source of information, particularly for the estimation of health risk exposures and outcomes. This dissertation studies how the same survey questions are answered differently depending on whether the respondent is a mother of a child in the 90s, the child as a young adult, or whether the respondent is female or male. Topics explored were chosen as those in which sex and social norms around sex and across generations can influence how we experience health, the risk factors we are exposed to, and our relationship with morbidity.

In the first chapter, *Differences in reporting of child abuse by mothers and young adults*, we used a longitudinal study to compare mothers' prospective accounts of their child experiencing different forms of violence against children (VAC), and young adults' retrospective self-reports of experiencing VAC. We then studied the socioeconomic factors associated with mothers reporting abuse of their child among children that were classified as having experienced abuse. For this end, we used the Avon Longitudinal Study of Parents and Children (ALSPAC), a 30-year prospective birth cohort study in England. This chapter addresses a well discussed issue in the field, the underreporting of VAC depending on the survey respondent. We used longitudinal data, unlike previous work where mostly cross-sectional data was used. We found that when questions are asked in the same way, there was no evidence of mothers underreporting physical or psychological abuse in comparison to children, even though there was little reliability across respondents. Among the pairs of mothers and children

in which at least one of them reported abuse, we found that the sex of the child and other mother characteristics are associated with mothers' reporting of physical or psychological abuse. Finally, the first chapter reflects on the social norms around discipline, as both mothers and the young adults described physical cruelty to be related to acts of severe physical violence, in which case acts such as pushing, smacking or kicking, would not be classified as abuse if mothers were the only respondents.

Chapters two and three focus on a fundamental topic in the measurement of the burden of disease, the measurement of morbidity through disability weights. In chapter two, *Differential health loss valuation by sex of the respondent in the Global Burden of Disease (GBD) study* we analyzed differences by sex of the respondents in the disability weights used for the GBD study. Similar to literature focused on paired comparison questions from Martens de Noordout et al. in 2018, Liu et al. in 2020, and Haagsma et al. in 2024, we found high correlation of health preferences by sex. This translated into anorexia nervosa being the only health state for which there a was a significant difference between disability weights estimated with female only and male only paired comparison data. In contrast, the sex stratification of the population health equivalence questions resulted in significantly different disability weights for females and males in almost all health states measured. In other words, we found that in the disability weights used in the GBD study, preferences for health states do not differ by sex, but females are less willing to accept disability as health program evaluations in comparison to males.

Finally, in chapter three, *Differential health loss valuation by sex on population health equivalence questions (PHE),* we further explored the sex differences found in chapter two and analyzed willingness to accept disability using the disability weights data from the GBD study. Through a marginal logistic model using generalized estimating equation, we found that even when we take age and education into account, females are more likely to choose programs that avoid deteriorating health over preventing death for the relative few. Before this study there was no empirical evidence on the differences in PHE valuation by sex of the respondent, mainly due to the use of PHE data as a methodological step. These questions are not used to rank health states but to anchor the preferences revealed through paired comparison questions in values that are useful for the estimation of Years Lived with Disability (0 to 1 ranges). It is in these questions, that we found females are more likely to choose the program that averts lifelong consequences of disease over programs that avert death as creating the greater population health benefit. Consequently, we estimated that

if all disability weights input data were stratified by sex, female disability weights would be larger for every heath state. Notably, because 70 percent of the respondents of population health equivalence questions are female, the current set of disability weights in the GBD study reflect more the preferences of disability weights of females than males.

Table of contents

Acknowledgments

I'm extremely thankful to my chair and my committee members for all their support throughout the research and writing of this dissertation. From the first steps of drafting the proposal to the final defense, I have always felt the support and constructive guidance of my committee. I am thankful for the Health Metrics Department, the Institute for Health Metrics and Evaluation and the disability weights experts I have been fortunate to interact with over the last years, thank you to professors Juanita Haagsma, Joshua Salomon, Shuhei Nomura, Yu Chuanhua, and Xiaoxue Liu. I have truly enjoyed working on disability weights and it often comes to my mind unprompted, brining me back to moments of reflection. Thank you Reed and Tom, it has been really fun and rewarding to explore disability weights together. I would also like to thank the Gender Equality Metrics team at IHME - Cory, Jack, Mariam, Molly, Carol, and the three Erins. I have enjoyed working with you and have learned a lot from your talent, determination and sense of responsibility. A special thank you to Emmanuela again, you have been a great advisor from the first day of the PhD program, I always felt extremely fortunate to have had the opportunity to learn from you. Applying to PhD programs was challenging and time-consuming, decades in the make, but throughout my time in the Health Metrics program I have always felt that I made the right decision, there is no other place I should have being studying and working at. Finally, thank you to my family, for absolutely everything.

Chapter 1: Differences in reporting of child abuse by mothers and young adults

Introduction

Overview

In this chapter we study differences in reporting child abuse from prospective accounts from mothers on behalf of their child, and retrospective self-reports from young adults. We first describe the current best practices for measuring violence against children (VAC), the debate around whether to use parental reports to estimate exposure, and the existent empirical evidence that focuses mainly on cross sectional studies. We then present the Avon Longitudinal Study of Parent and Children (ALSPAC) longitudinal data as a candidate to explore differences through generations, and dive into the methodology used to compare the mother's and the young adult's violence accounts of the young adult s childhood. Finally, we discuss the results from our analysis and reflect on the contributions made to the field of health metrics.

Prevalence and measurement of VAC

Violence against children is a major human right violation with immediate and long-lasting health, wellbeing, and socioeconomic consequences that affect a large percent of the population. Some accounts estimate 1 billion children experienced violence in the previous year[1]. More specifically, the estimated global age-standardized prevalence of sexual violence against children (SVAC) was 18.9 percent for females and 14.8 percent for males in 2023[2]. As a health risk factor, childhood abuse has been found to increase the risk for alcoholism, drug abuse, depression, suicide attempts, smoking, sexually transmitted diseases, physical inactivity, severe obesity, and becoming perpetrators and victims of Intimate Partner Violence later in life[3–15]. In recognition of this problem, the international community has established the clear goal of ending VAC by 2030[16], which makes the measurement of VAC a key component in monitoring progress.

Unfortunately, measuring VAC is challenging. The sensitive nature of the topic compounds the already known concerns of bias, accuracy and external validity health researchers' face when using different data sources to measure exposure to health risks and outcomes. In the field, underreporting

or no disclosure of VAC is well documented. Children fear the consequences of reporting, they lack the vocabulary to describe experiencing it, are subject to cultural stigmas and social norms that normalize VAC, and deal with distrust of authorities and inadequate reporting systems[17–19]. For many years, measurement of VAC was left to researchers and authorities from each individual country. The lack of a gold standard approach to collect and process VAC data made measurement of exposure subject to study level biases and idiosyncrasies[20]. It was not until 2023 that the International Classification of Violence Against Children (ICVAC) was released with clear definitions and examples of different forms of VAC[20].

ICVAC was a consequence of many efforts made over the last three decades to standardize the measurement of VAC, and therefore builds upon a body of literature that established best practices in estimating exposure[19,21]. Among these recommended best practices is the use of population-based survey data instead of administrative records of violence to provide more accurate prevalence estimates, the use of questions that ask respondents about specific acts instead of general abuse or violence, and the use of self-reported survey responses over parental reports on children s behalf. This last point being less prescriptive than the first two. The Global Burden of Disease study, for example, excludes studies where respondents are not answering for their own sexual violence experiences when estimating SVAC[22]. While UNICEF recommends using the Multiple Indicator Cluster Surveys to estimate VAC in which parents are asked about their use of violent discipline for children under 5 years old[23].

Theoretical concerns of self and parental reports of VAC

The discussion in the field on whether parents are reliable sources to assess VAC is not new. In 1998, Straus, Hamby and authors reviewed the consistency of parents' responses to child abuse and neglect in their adaption of the Conflict Tactic Scale. They found that the Parent-Conflict Tactics Scale (CTSPC) tool would work well to assess violence, but that researchers should consider prevalence from these estimates to be the minimum estimate of child maltreatment[24]. The concern being that using parental responses on population surveys would be biased given that most of the instances of child physical and psychological violence are perpetrated by parents or household members[25]. Social desirability bias would motivate parents to underreport their use of physical and psychological aggression when used to discipline their children. Moreover, shame and fear of legal consequences would drive parents to underreport violence[26]. Even in cases where violence is

perpetrated by someone outside of the household, children and teenagers may not disclose their experiences to their parents, making it impossible for parents to report it[26].

Retrospective self-reports from adults on their lived experiences of VAC have their particular and general set of limitations. Individuals can forget accounts from their childhood, timing of the experience can be inaccurate, and children could even not be aware certain experiences are violent while living through them, making recollection even more difficult, particularly if experiences of VAC happened before the child was three[27]. Aside from omission and retrieval biases, the nature of the exposure and questions to assert VAC is compounded due to stigma and social norms. Even if victims remember the experiences accurately, shame and fear of bringing up traumatic experiences also contribute to underreporting[28]. Although very few longitudinal studies are available to measure VAC, in a couple of studies where more than one round of VAC questions were asked from respondents, inconsistent self-reports of the experience of violence highlight the limitations of retrospectives accounts[29,30].

The nature of the exposure being measured makes it hard to rely on prospective self-report accounts of VAC. Although we are more likely to get closer to the true exposure of VAC when using prospective self-reports from children and adolescents, the ethical concerns and risks of increasing discomfort, distress or trauma from participants should not be taken lightly[31]. To minimize the probabilities of harming participants, data collection on VAC needs to follow context specific safeguarding plans[32]. In some contexts, these safeguards are not in place, and data obtained from the parents would be the only source of information that could be ethically used. It is therefore important to quantify the difference between accounts of violence depending on the respondents, and to understand the magnitude in the bias of VAC estimates depending on the respondent.

Empirical research on self and parental reports of VAC

Surprisingly, not much empirical evidence quantifies the difference in VAC estimates from children self-reports and their parents. Chan 2015 used a matched sample of parents and their children in a Chinese cross-sectional survey from 2009 that used the Juvenile Victimization Questionnaire to assess a broad definition of child victimization. Chan found low levels of agreement for all forms of victimization (e.g. including child maltreatment), and significantly higher prevalence of lifetime and last year experience of victimization types reported by adolescents compared to parents[26]. Sofuoglu Z. et al. 2015 assessed physical and psychological maltreatment in a sample of 2,608 pairs of parents

and children in Turkey using the International Society for the Prevention of Child Abuse and Neglect surveys to parents and children (e.g. ICAST-P and ICAST-C) and found differences in the proportion of exposure depending on the respondent, with parents inclined to underreport[33]. Sierau S. et al. 2018 used a modified version of the Parent-Child Conflict Tactics Scales to ask children about their experience of violence in a sample of 811 pairs of children and their parents in Germany, and found low correlation between the assessment of violence for physical, emotional and sexual abuse[34]. In 2024, Hogan et al. used one round of the US Adolescent Brain Cognitive Development study to compare children and their parents responses on whether family members sometimes hit each other, they found that of those reporting physical violence the majority were not concordant responses between parents and children, with children again reporting higher exposure[35].

There is one study that investigates the concordance between parents and children responses in a longitudinal study. In 2024, Dunn E. et al. used data from ALSPAC, a prospective birth cohort study in the United Kingdom, to compare prospective parental reports of physical and emotional child violence perpetrated by the parents to the retrospective accounts of young adults of physical and emotional abuse perpetrated by an adult in the family, and the factors associated with disagreement in reporting[36]. They found low levels of agreement between parent and young adult reports. Although an important contribution to the literature, a detailed review of the ALSPAC study reveals missed opportunities in the analysis of this data. First, sexual VAC is omitted from the study. Second, even though prospective accounts of parental report of VAC exist until the child is 18 years of age, Dunn et al. only compared experience of violence when the child is up to 9 years of age. Third, the questions being compared between parents and young adults are not the same. In the parents survey, the question asks broadly whether the parents were physically and emotionally cruel towards the children. In the young adult questionnaire, the 22 year old respondent is asked whether they experienced different behaviors that can be categorized as physical and emotional violence by an adult in the family. Dunn et al., omitted questions when the young adult is 23 and 27 years old, in which they report on their VAC experiences. This last point is particularly important given that the question at 27 years of age corresponds exactly to the questions asked from parents: Before you were 19 a parent was physically/emotionally cruel to you?". Finally, when studying the factors associated with parental and child disagreement in reporting, Dunn et al restrict their analysis to univariate regressions given the large number of missing values in the data.

Purpose of this study

Given the lack of data in the measurement of VAC, the limited empirical literature comparing parental and children reports from longitudinal studies, and the richness of the ALSPAC data to analyze in detail the differences between mothers and young adults reports of VAC, we used the ALSPAC data to 1) compare mothers' prospective accounts of her child experiencing different forms of VAC, and young adults' retrospective self-reports of experiencing VAC, and 2) study the socioeconomic factors that are associated with mothers reporting abuse of their child among children that experienced abuse.

Methods

Data

We used data from ALSPAC, a population-based prospective birth cohort study[37,38]. The objective of the study was to understand how an individual s genotype, combined with environment, determines health and development. Pregnant women residing in Avon, UK with expected dates of delivery between 1st of April 1991 and 31st of December 1992 were invited to take part in the study. Initially, 14,541 pregnancies were enrolled. Of these, 13,988 children were alive at 1 year of age. Additional recruitment efforts when the child is seven years old increased the total sample size to 15,447 pregnancies, of which 14,901 children were alive at 1 year of age[39]. Similarly, because of additional recruitment efforts, as of September of 2021, 14,833 unique women (G0 mothers) were enrolled in ALSPAC. Although recruitment was done with the intention of being representative of the Great Britain population, ALSPAC mothers were found to be more likely to live in owner-occupied accommodations, to have a car, and to be white in comparison to Avon and Great Britain[37,38].

ALSPAC is a very large and complex study, with different questions asked to mothers, partners, and children. Data were collected from a variety of sources such as self-completion questionnaires, medical and education records, ad hoc visits to participants households, biological samples, and in-depth interviews to smaller samples in order to assess the validity of the survey estimates[40]. In this study, we focused solely on data collected by self-completion questionnaires given to mothers and their offspring or focus child. Going forward the focus child will also be referred to as the young adult given that they completed the questionnaires of interest at 22 and 27 years of age.

Violence against children definition

Two frameworks informed the definition of VAC in this study. The first one is Krug et al. 2002 typology of violence, which classifies violence into three categories: self-directed violence, interpersonal violence, and collective violence[41]. This framework also distinguishes between the nature of violence as physical, sexual, psychological, and deprivation or neglect. The second framework is ICVAC, which classifies specific behaviors into different types of violence again as physical, sexual, psychological, and deprivation or neglect[20]. Our research focused mainly on family interpersonal violence, which occurs in the house of the children, and where the perpetrators are family adults or the parents with an exception for sexual violence, in which case any adult perpetrator is considered. Given the power dynamic between perpetrator and child for physical and psychological violence, our study is referring to child abuse for these forms of violence. On the contrary, because we do not have the information of the perpetrator of sexual violence, when discussing sexual violence against children we are not exclusively referring to abuse. However, for ease of expression, we use the term abuse interchangeably with VAC in the rest of the chapter.

To define child abuse in our study, we thoroughly reviewed all violence related questions in all self-completed ALSPAC questionnaires, crosschecked them with the literature that used VAC or Intimate Partner Violence (IPV) from ALSPAC as either exposure or health outcomes, and compared them to the questions in common population-based survey instruments measuring VAC. Appendices 1 through 4 present the survey instruments reviewed as well as the ALSPAC violence questions identified. We found a variety of ways that ALSPAC asked about child abuse. Most of the questions on whether the focus child experienced physical or psychological violence in the household were asked to mothers and partners in the same way, generally and repeatedly over time. Mothers and their partners were asked whether they (and their partners) were physically or emotionally cruel to the focus child since the last time they were interviewed. These were prospective questions asked eleven times before the focus child turned 18-years-old. For sexual violence, all prospective questions were asked to the mother and did not specify the identity of the perpetrator. The young adult responded for themselves retrospectively about violence in interviews when they were 22, 23 and 27 years old. At 27, the young adult received the same general question about their experience of violence, while at 22 they received more specific behavioral questions. We decided not to use questions asked when the young adult is 23-years-old given how different they are in comparison to questions asked to the mother and their little to no use in the literature (Appendices 2 to 4).

Table 1 presents the questions we used in this study to define familial interpersonal violence against the young adult. We selected the violence questions that could be most comparable between respondents. For example, although the ICVAC definition of physical and psychological violence were not restricted to the perpetrators being a parent of the child, we only used psychological and physical questions in which the parent was the perpetrator. In contrast, for child sexual abuse any person could be the perpetrator as this was not specified in the questions.

Table 1. Definition of child abuse by respondent

Type	**Respondent**	**Question**	**Definition**	**Age of child at round**
Physical	Child	Before the age of 11/between the ages of 11 and 17, how often did an adult in your family (anyone you consider to be a family member): c. Push, grab or shove you d. Smack you for discipline g. Actually kick, punch, or hit you with something that could hurt you or physically attack you in another way h. Hit you so hard it left you with bruises or marks	Yes if any of these behaviors were experienced by respondent	22 years
	Child	Before the age of 11/between the ages of 11 and 17, how often did an adult in your family (anyone you consider to be a family member): h. Hit you so hard it left you with bruises or marks	**Severe physical**: Yes if any of these behaviors were experienced by respondent	22 years
	Child	Before you were 19 a parent was physically cruel to you	Yes if respondent did not select option " No, did not happen"	27 years
	Mother	Since [age of child at last interview or time that has passed since last interview]: You/your partner was physically cruel to your children?	Yes if respondents answered yes for her or her partner in any of the rounds	0.6, 1.75, 2.75, 3.9, 5, 6, 9, 11, 18 years

Type	Respondent	Question	Definition	Age of child at round
Psychological	Child	Before the age of 11/between the ages of 11 and 17, how often did an adult in your family (anyone you consider to be a family member): a. Shout at you b. Say hurtful or insulting things to you e. Punish you in a way that seemed cruel f. Threaten to kick, punch, or hit you with something that could hurt you or physically attack you in another way	Yes if respondent selected either often or very often for any of the behaviors	22 years
	Child	Before you were 19 a parent was emotionally cruel to you	Yes if respondent did not select option " No, did not happen"	27 years
	Mother	Since [age of child at last interview or time that has passed since last interview]: You/Your partner was emotionally cruel to your children?	Yes if respondents answered yes for her or her partner in any of the rounds	0.6, 1.75, 2.75, 3.9, 5, 6, 9, 11, 18 years
Sexual	Child	Before the age of 11, were you touched in a sexual way by an adult or an older child or were you forced to touch an adult or older child in a sexual way when you did not want to?	Yes if this happened more than once for any of the two questions	22 years
	Child	Before the age of 11, did an adult or an older child force you or attempt to force you into any sexual activity by threatening you or holding you down or hurting you in some way when you did not want to?		
	Child	Before you were 19 you were sexually abused	Yes if happened under 11 years old	27 years
	Mother	Since [age of child at last interview or time that has passed since last interview] state whether any of these happened: She was sexually abused	Yes if respondents answered yes in any of the rounds	1.5, 2.5, 3.5, 4.75, 5.75, 6.75, 8.6 years

Study data after the child was 22 years old were collected and managed using REDCap electronic data capture tools hosted at the University of Bristol[11]. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies. The study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/).

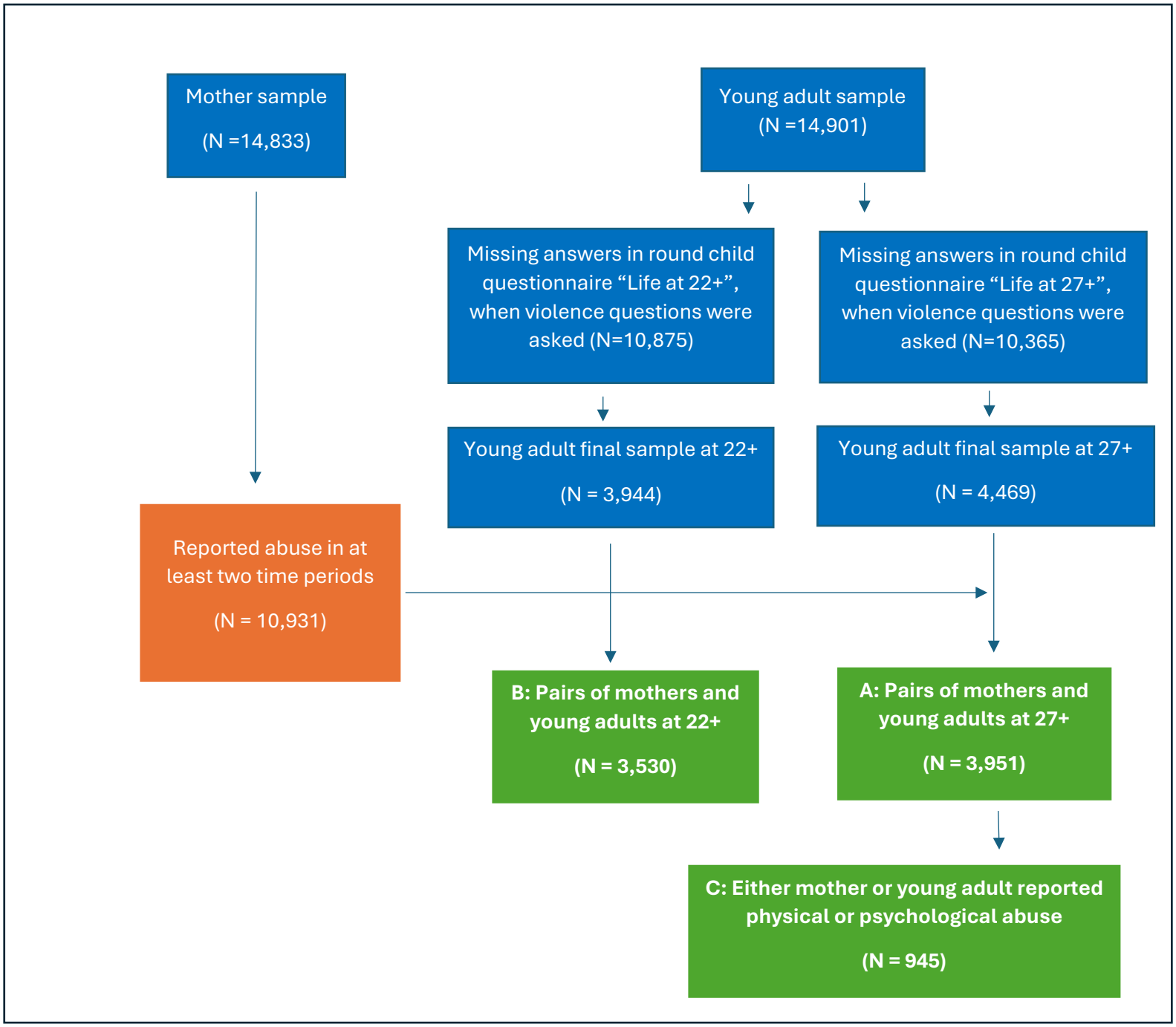
We defined sexual and physical abuse as having occurred if respondents said yes to any questions no matter the frequency or whether the respondent was the mother or young adult. When the young adult is the respondent, exposure to psychological abuse was defined only if respondent said at least

one of the acts happened often or very often. When the mother is the respondent, exposure to different types of abuse are defined if she replied yes to the question without any reference to frequency. For the young adult, the questions are retrospective and ask about lifetime child violence. For the mother, the questions are prospective and ask respondents to think about the period since the last interview. When comparing mothers' and their children self-report of child abuse, we use a cumulative prevalence of violence. Meaning, a child was classified as experiencing abuse if mothers' reports of violence on behalf of their child was yes in any of the rounds where she was asked the violence questions.

Sample size and inclusion criteria

Characteristic of longitudinal studies that span decades, the ALSPAC study has a lot of missing values. When the young adult is 22 years old the sample of index children drops to 3,944, 26.5 percent of the initial observations received and to 4,469, when the young adult is 27 years old, 30 percent of the initial observations. Figure 1 describes the number of individuals lost to attrition as well as the samples used for analysis.

Figure 1. Analysis sample sizes



We used two samples to compare mother and young adult self-reports of child abuse. In sample A of Figure 1 we kept the pairs of young adults that answered questions of violence in round Life at 27+", and of mothers who responded child abuse questions prospectively in at least two rounds. This was the primary sample used to compare abuse by respondents given that abuse classification

depends on the same questions for both mothers and young adults: “were parents physically or emotionally cruel to children?”, and “was the child sexually abused?”. We also analyzed differences in the exposure of violence of the young adult by respondent with sample B in Figure 1. Sample B was restricted to the pairs of mothers who responded to at least two rounds of violence questions about the young adult, and the young adults who responded to violence questions at 22 years old. Finally, sample C was used to investigate the socioeconomic factors associated with mothers reporting abuse among young adults that experienced VAC. Sample C was therefore restricted to the pairs of respondents in which either the mother or the young adult reported child abuse.

Partners of the ALSPAC mothers were also asked the same questions on physical and psychological abuse of their child. However, we decided to exclude these questions from the analysis because the pairs of non-missing observations for sample A would have left us with 1,513 observations. Moreover, Dunn et al. 2024 already explored this comparison and found high concordance between mothers and their partners responses on their child experiencing physical or psychological abuse[36].

Statistical analysis

We estimated the proportion of young adults that experienced physical, psychological and sexual child abuse from samples A and B in Figure 1, and used Wilson score interval method to estimate the standard errors[42]. We then checked if differences in proportions were significant by estimating Fisher test statistics. We also tested concordance of the responses between young adults and mothers by estimating Cohen s kappa and Prevalence-Adjusted Bias-Adjusted Kappa (PABAK)[43,44]. Although Cohen s kappa is broadly used to check the concordance when studying the psychometric properties of survey instruments, PABAK is more appropriate for rare outcomes[43].

To identify the socioeconomic factors that were associated with mothers reporting abuse on children that experienced abuse (as defined by either the young adult self-report of VAC or the mother reporting abuse of the focus child), we ran three multivariate logistic regressions with the following specification:

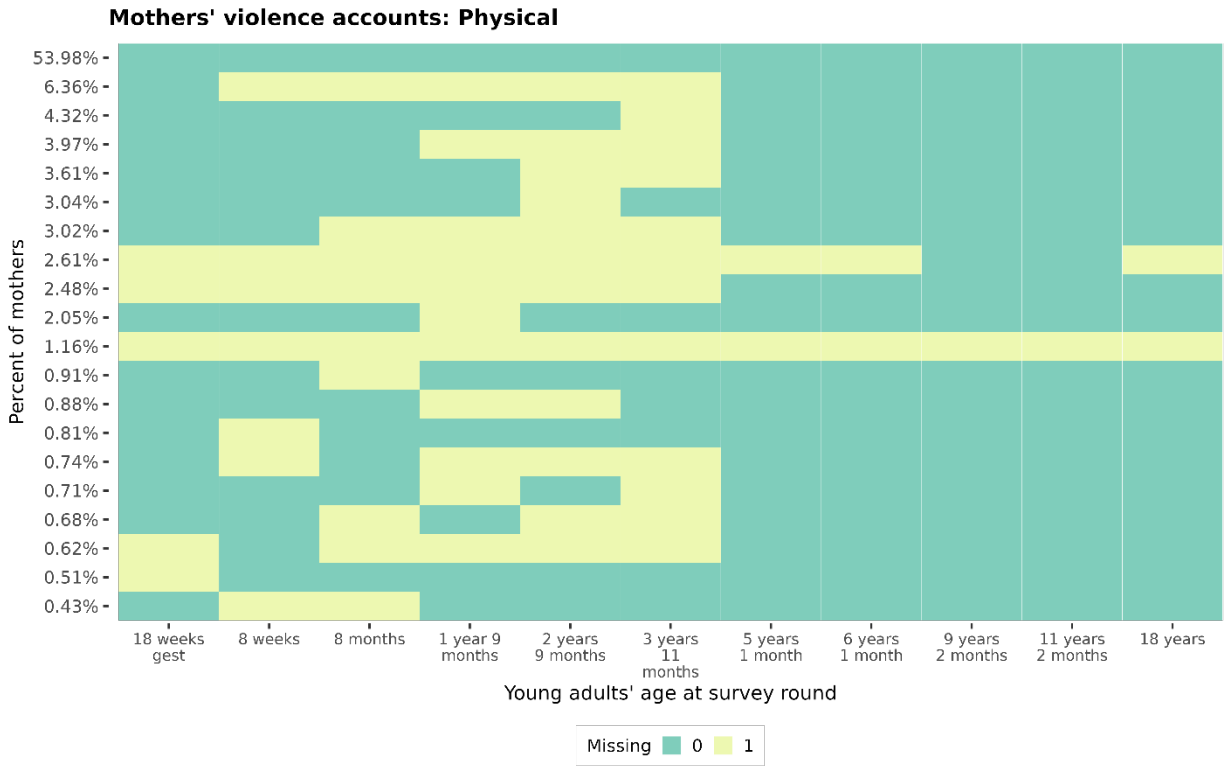
$$logit\ \Pr(Y_{i,} | X_i) = \beta_0 + X_i\beta \qquad (1)$$

Where $Y_{i,}$ is a binary variable indicating whether the mother reported her child experienced either any physical or psychological, physical, or psychological abuse from her or her partner. For all three outcomes, the covariates of interest (X_i) were selected socioeconomic variables similar to what

has been used in previous literature that studied the parental factors associated with disagreement in reporting VAC[26,35,36,45]. We used the young adult s sex and race, and mother s age, number of other children, education, civil status, and socioeconomic class at pregnancy as covariates. We also included mother s own experience of child abuse, whether she reported experiencing physical, psychological, sexual abuse, neglect and whether she witnessed any parental IPV when growing up. Appendix 5 describes the variables used for mother s own experience of child abuse and neglect.

We observed different types of missing data throughout the ALSPAC study. Some individuals missed some rounds but returned, others left the study after a number of rounds, and there were also item nonresponses. Figure 2 shows a non-monotone pattern of missing data for physical abuse as reported by the mothers throughout time. Young adults and mothers experiencing economic hardship were more likely to skip questionnaires or leave the study altogether[46]. We did not impute missing values for our outcomes. The sample at round Life at 27+" is 26.5 percent of the initial 14,901, any imputation model that attempted to impute 74 percent of the data would have required unrealistic assumptions. Instead, we acknowledged the reduction in sample as a limitation of our study, and used multiple imputation models using chained equations (MICE) to impute only values of predictors in our multivariate logistic regression[47]. The variables used in the imputation model were based on Houtepen et al. 2018 and Chan et al. 2021 studies which used this method to estimate adverse childhood experiences from ALSPAC, and the effect of maternal IPV and financial adversity on homelessness[48,49]. Appendix 6 presents all variables for the imputation.

Figure 2. Pattern of missingness for child physical abuse in mothers' reports



Note : Missing pattern for 100 percent of observations are not presented as the plot would have been too long.

We ran MICE with fully conditional specification (FCS) in R. Among the predictors for the logistic regression, 34 percent of the observations had at least one missing variable. Following Von Hippel s rule of thumb of imputation per percentage of incomplete cases[50], we ran 35 imputations with 50 iterations. We used predictive mean matching to impute continuous variables, logistic regression for binary variables, and polytomous logistic regression for categorical variables. We derived final estimates from the imputed model by combining within-imputation variance and between-imputation variance using Rubin s rules to account for uncertainty due to missing data[47,51].

We ran several sensitivity analysis. First, we compared our imputed model estimates to the complete case or individuals with complete answers logistic regression. Second, we used both young adult's responses in rounds "Life at 22+" and "Life at 27+", combined the responses and estimated a variable that indicated young adult reporting any type of abuse in any of the two rounds. Third, we compared the proportion of young adults experiencing different types of child abuse when we restrict the sample to pairs of mothers and young adults in which mothers responded to all of the questions about violence towards their child. For all three comparisons, we imputed missing predictors using

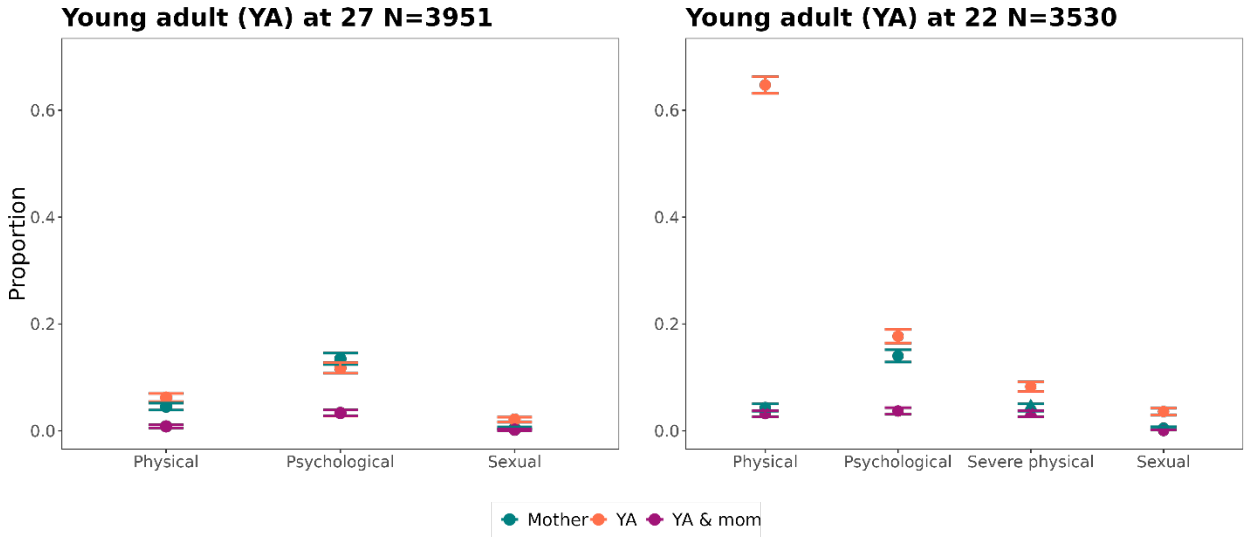
MICE, and then restricted our sample to observations where either the young adult or the mother reported abuse for the logistic regression analysis.

Results

Comparison of mothers' and young adults' report of violence

Figure 3 shows the proportions of children who reported having experienced different forms of VAC. In the left plot we observe the pairs of mothers and young adults where retrospective self-report of abuse was taken at 27 years of age, in the right plot we observe the pairs of mothers and young adults' retrospective VAC reports at 22 years of age. When we compared the proportions of young adults who experienced VAC by respondent, we found that physical cruelty" was interpreted by mothers as severe physical violence acts. The proportion of young adults at 22 years old that reported severe physical child abuse, 0.08 (95% CI 0.07-0.09), is much lower than any child physical abuse 0.65 (95% CI 0.63-0.66), and closer to the proportion of mothers that reported either her or her partner being physically cruel to the child, 0.04 (95% CI 0.04-0.05). The difference between the young adults' report of physical abuse and the mothers' report of abuse, as well as the little difference between the young adults' report of severe physical abuse and mothers' report of abuse made us conclude that when mothers thought about "cruelty" they were recalling more severe acts of physical violence: hit so hard that were left with bruises or marks" as described in Table 1. Interestingly, the proportion of young adults that reported experiencing physical cruelty as children by any of their parents was 0.06 (95%CI 0.05-0.07), much closer to the mothers' report of 0.05 (0.04-0.05) (left panel in Figure 3). Again, it appears young adults also interpreted cruelty" as severe acts of physical violence. In other words, neither young adults or their mothers considered acts such as pushing, grabbing, shoving, smacking for discipline, kicking, or punching to be cruel.

Figure 3. Proportion of young adults experiencing abuse by respondent and survey round



Note: In the right panel, there are two accounts of physical abuse, severe and physical. If a young adult reported being hit so hard that he/she was left with bruises or marks before the age of 18, this was classified as severe abuse. Mothers were only asked about one type of physical abuse whether she or partner were physically cruel to the child".

Although the proportion of mothers and young adults that reported violence were similar, these estimates were statistically different (Table 2). Fisher exact test p-values were less than 0.05, meaning we rejected the null hypothesis of association between mothers and young adults' reports for all forms of violence both when young adults answered retrospectively at 22 or at 27 years old. Except for sexual abuse in round Life at 22+", for which the sample is so small that we were restrained from drawing any conclusions for this estimate. Table 2 also shows low levels of agreement between responses when we use Cohen s kappa estimates. Cohen s kappa estimates were 0.11, 0.16, and 0.09, for physical, psychological and sexual abuse when the young adult responded at 27 years old. When we analyzed agreement accounting for the low prevalence of our outcomes, PABAK estimates showed higher agreement: 0.82, 0.63 and 0.96 respectively. In other words, there was high agreement between young adults and mothers on young adults not experiencing child abuse, but low agreement when violence was reported. The mothers and young adults who reported abuse are not the same, which makes the proportion of young adult and mother pairs that reported abuse considerably smaller than only mothers and only young adults in Figure 3.

Table 2. Proportions of physical, psychological and sexual abuse per respondent

		Young Adult report		Mother Report		Young Adult vs. mother		
Sample	**Form of abuse**	**M**	**Proportion (95% CI)**	**M**	**Proportion (95% CI)**	**Cohen's Kappa (95% CI)**	**PABAK (95% CI)**	**Fisher (p-value)**
At 27	Physical	0	0.06 (0.05-0.07)	0	0.05 (0.04-0.05)	0.11 (0.02-0.2)	0.82 (0.8-0.84)	3.77 (0)
At 27	Psychological	0	0.12 (0.11-0.13)	0	0.13 (0.12-0.15)	0.16 (0.1-0.21)	0.63 (0.6-0.65)	3.06 (0)
At 27	Sexual	9	0.02 (0.02-0.03)	0	0.01 (0-0.01)	0.09 (-0.09-0.27)	0.95 (0.94-0.96)	16.38 (0)
At 22	Physical	0	0.65 (0.63-0.66)	0	0.04 (0.04-0.05)	0.01 (-0.01-0.04)	-0.25 (-0.28--0.22)	1.58 (0.02)
At 22	Severe physical	0	0.08 (0.07-0.09)	0	0.04 (0.04-0.05)	0.08 (-0.01-0.16)	0.78 (0.76-0.8)	2.74 (0)
At 22	Sexual	40	0.04 (0.03-0.04)	124	0.01 (0-0.01)	0.02 (-0.14-0.18)	0.92 (0.91-0.93)	3.6 (0.12)
At 22	Psychological	0	0.18 (0.16-0.19)	0	0.14 (0.13-0.15)	0.09 (0.04-0.14)	0.51 (0.48-0.54)	1.85 (0)

Note: M= Missing. In sample "At 22", mothers; report of physical violence is the same for physical and severe physical, it is being compared against two different definitions of physical abuse from the index child questionnaire.

Sexual abuse prevalence was so small among mothers' reports that we did not further analyze the factors associated with mothers reporting this form of abuse and only focused on physical and psychological abuse.

Socioeconomic factors associated with mothers reporting child abuse among young adults that were classified as experiencing child abuse

Table 3 shows the descriptive statistics for the pairs of mothers that reported prospectively on child abuse and young adults that reported retrospectively on experiencing child abuse at round "Life at 27+". The questions used to assert abuse here were the same. Table 3 shows the percentage of missing values for different socioeconomic predictors for both the complete case sample of mother and young adult pairs, and the imputed data. A complete case analysis would have resulted in 34 percent of the paired observations being dropped. The auxiliar variables used for imputation are described in Appendix 6. Means and proportions of the predictors are very similar between complete cases and multiple imputations. Of the 3,951 pairs of young adults and mothers in our sample, 65 percent were female and 96 percent white. The majority of mothers had other children, and were in their thirties when pregnant with the young adult. Almost half of the mothers had more than a high school education and almost all of them were part of social class I to III (e.g. Professionals, managerial and technical, and skilled non manual) as described by the 1991 UK Office of Population Census and Survey classification[48]. A small percentage of mothers experienced psychological and physical abuse themselves, seven and four percent respectively. On the contrary, around a quarter of the mothers in our sample experienced sexual abuse. This is not surprising as it can be seen in

Appendix 5 that for mothers' experience of child sexual abuse we used specific questions that asked about behaviors, in contrast to a single general question on both physical and psychological child abuse.

Table 3. Proportions of physical, psychological and sexual abuse per respondent in Life at 27+ sample

			Complete cases		Imputed+complete cases	
Child predictors	N	Percent Missing	Mean or Proportion	Standard Error	Mean or Proportion	Standard Error
Female child	3951	0.000	0.645	0.008	0.645	0.008
White child	3804	3.720	0.965	0.003	0.964	0.003
Mother predictors at pregnancy						
Age	3900	1.290	28.9	0.072	28.9	0.071
N other children	3831	3.040	0.740	0.014	0.742	0.014
Married	3901	1.270	0.837	0.006	0.835	0.006
Education						
More than high school	3863	2.23	0.474	0.008	0.470	0.008
High school	3863	2.23	0.346	0.008	0.349	0.008
No high school	3863	2.23	0.180	0.006	0.181	0.006
Social class						
I & II	3384	14.350	0.451	0.009	0.432	0.008
III	3384	14.350	0.469	0.009	0.482	0.008
IV & V	3384	14.350	0.079	0.005	0.086	0.004
Psychological child abuse	3851	2.530	0.067	0.004	0.068	0.004
Physical child abuse	3926	0.630	0.041	0.003	0.041	0.003
Neglect	3627	8.200	0.202	0.007	0.201	0.006
Witnessed parental IPV	3463	12.350	0.113	0.005	0.111	0.005
Sexual child abuse	3580	9.390	0.251	0.007	0.252	0.007

We ran three logistic regressions on the imputed datasets and pooled the coefficients together in Table 4: any (physical or psychological), physical, and psychological abuse. The samples were small because we were interested in the mother s socioeconomic factors that are associated with mothers reporting abuse of their child when either her or the young adult reported abuse. Although we did not know the true prevalence of abuse, we assumed that either one of the respondents indicating abuse is enough to classify a young adult as being a victim of child abuse. Sex of the child was statistically significantly associated with mother reporting violence. Mothers reporting physical or psychological abuse towards their child were 0.69 (95% CI 0.51-0.93) less likely to do so if the child was female in in comparison to male. Moreover, the odds ratio for sex of the child was significant for psychological

abuse but not for physical abuse at the 5 percent level. Interestingly, mother's own experience of child abuse in comparison to not experiencing abuse is significant at the 5 percent level; mother s own experience of physical, and witnessing IPV increases the odds of mother reporting by 2.16 (95%CI 1.12-4.16), and 1.59 (95%CI 1.02-2.48) respectively. We also found that mothers with more than one child were more likely to report them or their partners being cruel to the index child, with each additional child being associated with an odds ratio of 1.26 (95%CI 1.05-1.51) in comparison to no other children. Although we do not have enough power to find any other significant associations, the directions of our estimated odds ratios coefficients signal that older mothers, with higher levels of education could have been more likely to report abuse.

Table 4. Multivariate logistic regression on mothers reporting of abuse with imputed data

	Any		**Physical**		**Psychological**	
	OR (95% CI)	**P values**	**OR (95% CI)**	**P values**	**OR (95% CI)**	**P values**
Child predictors						
Female child	0.69 (0.51-0.93)	0.01	0.64 (0.41-1.00)	0.05	0.65 (0.48-0.89)	0.01
White child	1.23 (0.65-2.35)	0.52	1.95 (0.76-5.04)	0.17	0.84 (0.41-1.72)	0.64
Mother predictors at pregnancy						
Age	1.03 (0.99-1.07)	0.10	1.10 (1.04-1.16)	0.00	1.04 (1.00-1.07)	0.05
N other children	1.26 (1.06-1.51)	0.01	1.19 (0.92-1.55)	0.19	1.28 (1.06-1.54)	0.01
Education						
High school	0.75 (0.52-1.07)	0.11	0.69 (0.39-1.22)	0.20	0.78 (0.54-1.13)	0.19
No high school	0.76 (0.50-1.16)	0.20	0.43 (0.22-0.87)	0.02	0.72 (0.46-1.13)	0.15
Married	1.19 (0.85-1.67)	0.31	1.38 (0.81-2.34)	0.24	1.17 (0.83-1.67)	0.38
Social class						
III	0.84 (0.60-1.19)	0.34	1.05 (0.60-1.85)	0.85	0.87 (0.60-1.26)	0.47
IV & V	1.00 (0.57-1.77)	1.00	1.71 (0.72-4.10)	0.23	1.03 (0.57-1.86)	0.92
Psychological child abuse	0.83 (0.50-1.36)	0.46	0.64 (0.30-1.38)	0.25	0.85 (0.51-1.41)	0.52
Physical child abuse	2.16 (1.12-4.16)	0.02	3.25 (1.38-7.66)	0.01	2.02 (1.03-3.95)	0.04
Neglect	1.00 (0.69-1.44)	1.00	0.85 (0.46-1.57)	0.61	1.05 (0.72-1.53)	0.81
Witnessed parental IPV	1.59 (1.02-2.48)	0.04	1.33 (0.68-2.60)	0.41	1.63 (1.03-2.59)	0.04
Sexual child abuse	1.34 (0.99-1.82)	0.06	1.24 (0.77-1.99)	0.38	1.31 (0.95-1.82)	0.10
N	945		392		865	
AIC	1226.19		519.59		1125.33	

Our results are robust through multiple sample specifications. Appendix 7 compares the imputed and complete cases models, the direction and magnitude of the odds ratio are similar, but the complete case analysis lacks the power to identify statistically significant associations. Appendix 8 presents the results when the sample being considered is whether the young adult reported experiencing child abuse in either of the rounds at "Life at 22+", and "Life at 27+", the direction of the

odds ratio and magnitude is consistent with Table 4. Finally, in Appendices 9, 10 and 11 we compared the responses between mothers and young adults when we restrict the mothers' sample to only those who were not missing any of the prospective VAC questions, again the results are consistent with our main findings.

Discussion

Measuring exposure to VAC is difficult, it requires an understanding of all the different sources of biases pertaining to different data collection methods, and their interaction with the nature of the exposure. In this chapter we tackle one of the sources of biases in population-based surveys, whether to trust parental responses on VAC when they or their partners could be the perpetrators. Through our analysis of a longitudinal cohort study, we found that when questions use the same phrasing, there is little difference between the prevalence of child physical and psychological violence as reported prospectively by the mother of the child, and as retrospectively reported by the child as a young adult. However, we also find low to medium levels of agreement between respondents; that is, there is high agreement between pairs of mothers and young adults on not experiencing child abuse, but those young adults and mothers that report abuse are not exactly the same. We therefore conclude that the measurement of physical and psychological VAC could benefit from parental reports of violence even when there are no responses available from the child or young adult. Even with medium levels of agreement, the higher levels of parental reports of physical and psychological abuse in our comparison suggest that there is lower risk of parental underreporting for these forms of child abuse than for sexual child abuse.

We acknowledge that asking broadly about whether a parent was physically or emotionally cruel to their child lacks specificity and makes interpretation of cruelty subjective. However, given the different ways in which ALSPAC asked about VAC, we were able to compare young adults' general and specific report of physical violence to mothers' general reports of violence. We found that for both mothers and young adults, physical cruelty seemed to refer to severe acts of physical child abuse. Previous VAC literature using ALSPAC did not use young adults' questions when they are 27 years old, they focused mainly on the more specific questions asked when the young adult is 22 years old (Appendices 2 to 4). Our findings suggest there is more to be gained in the analysis of VAC in ALSPAC by also adding the young adults general questions on their experience of cruelty as children.

Finally, we find that given any report of physical or psychological abuse by the mother or the child, the child being a female is associated with a lower likelihood of the mother reporting her child experiencing any type of cruelty. This is particularly interesting when we take into consideration that 65 percent of the young adults in the logistic regression sample are females. In other words, even when there are more females that experienced abuse, mothers are less likely to report abuse perpetrated to their female child. This finding is driven by psychological cruelty, as there is no statistical power for a significant association between sex of the respondent and the mother reporting physical cruelty toward her child. Likewise, mothers who experienced physical child abuse themselves, and who witnessed parental IPV as children are more likely to report them or their partners perpetrating physical or psychological abuse.

An important limitation in the use of ALSPAC data is the large attrition, this is not uncharacteristic of longitudinal studies particularly when spanning more than 30 years. However, given the purpose of this study, we limited our sample size to the number of children that did not leave the study, similar to Yakubovich et al. 2019, Herbert et al. 2021 and Herbert et al. 2023[52–54] were they analyze the risk factors associated with the young adults' experiences with IPV. Attrition makes it difficult to claim exposure of VAC estimated in this analysis as representative of the general UK population. However, this was an ALSPAC limitation since its first rounds. Previous studies on ALSPAC reported the ALSPAC sample being more affluent and non-representative of the Avon region or the UK[37,38]. Overall, we argue that the benefit of studying differences in reporting of child abuse by respondents in a longitudinal study are far greater than the limitations due to attrition. Unlike cross sectional studies, in which parents are asked about perpetrating violence at one point in time, we were able to use mothers' reports over time which allow us to compare exposure over the entirety of childhood, meaning until the child is 18 years old. Moreover, our estimates of child physical and psychological abuse as self-reported by the young adults in the ALSPAC study are not that different to UK national estimates. In their 2023 study on prevalence of VAC in the United Kingdom, Nation et al. estimated physical child abuse prevalence in the UK to be 7.30 percent (95% CI 1.38-14.35), compared to our estimate of 6.22 percent (95%CI 5.47-6.98), and psychological child abuse to be 11.84 (95%CI 5.58-19.89) compared to our estimate of 11.77 percent (95% CI 10.76-12.78)[55].

A second limitation is our restriction of child sexual abuse to experiences reported by 11 years of age. Mothers only report prospectively on this question up to when the child is around 9 years of age, while young adults reported any sexual experience before 11 years old. As described by Cagney et al., most

experiences of sexual violence against children occur from 13 to 18 years old[2], which could partly explain the low proportions of child sexual abuse we estimated and limit our ability to make conclusions on sexual violence against children in this study.

This research provides further insight into the measurement of violence against children and makes a case for incorporating parental accounts of perpetration of physical and psychological violence in the estimation of its prevalence. Although the levels of agreement were low to moderate in our sample, we did not find that the mothers' cumulative prospective responses were underreporting these forms of abuse in comparison to the young adults' retrospective responses. In other words, we provide empirical evidence that physical and psychological forms of violence would be deemed less stigmatizing than sexual violence, and to conform more closely to social norms around discipline. Future data collection on this topic should encourage the use of more standardized VAC questions that specify acts instead of general violence questions in longitudinal studies so both the reliability within and across respondents can be further tested. More research is needed particularly to understand how cultural norms around discipline in different geographical contexts can affect differential reporting of VAC by parents and children.

Ethical approval, informed consent, and acknowledgments

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Bibliography

1 Hillis S, Mercy J, Amobi A, Kress H. Global Prevalence of Past-year Violence Against Children: A Systematic Review and Minimum Estimates. *Pediatrics* 2016; **137**: e20154079.

2 Cagney J, Spencer C, Flor L, *et al.* Prevalence of sexual violence against children and age at first exposure: a global analysis by location, age, and sex (1990–2023). *The Lancet* 2025; **0**. DOI:10.1016/S0140-6736(25)00311-3.

3 Felitti VJ, Anda RF, Nordenberg D, *et al*. Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study. *Am J Prev Med* 1998; **14**: 245–58.

4 Gilbert R, Widom CS, Browne K, Fergusson D, Webb E, Janson S. Burden and consequences of child maltreatment in high-income countries. *The Lancet* 2009; **373**: 68–81.

5 Walker HE, Wamser-Nanney R. Revictimization Risk Factors Following Childhood Maltreatment: A Literature Review. *Trauma Violence Abuse* 2023; **24**: 2319–32.

6 Fulu E, Miedema S, Roselli T, *et al*. Pathways between childhood trauma, intimate partner violence, and harsh parenting: findings from the UN Multi-country Study on Men and Violence in Asia and the Pacific. *Lancet Glob Health* 2017; **5**: e512–22.

7 Ehrensaft MK, Cohen P, Brown J, Smailes E, Chen H, Johnson JG. Intergenerational transmission of partner violence: a 20-year prospective study. *J Consult Clin Psychol* 2003; **71**: 741–53.

8 Abramsky T, Watts CH, Garcia-Moreno C, *et al*. What factors are associated with recent intimate partner violence? findings from the WHO multi-country study on women's health and domestic violence. *BMC Public Health* 2011; **11**: 109–109.

9 Anda RF, Felitti VJ, Bremner JD, *et al*. The enduring effects of abuse and related adverse experiences in childhood. *Eur Arch Psychiatry Clin Neurosci* 2006; **256**: 174–86.

10 Miller E, Breslau J, Chung W-JJ, Green JG, McLaughlin KA, Kessler RC. Adverse childhood experiences and risk of physical violence in adolescent dating relationships. *J Epidemiol Community Health* 2011; **65**: 1006–13.

11 Fonseka RW, Minnis AM, Gomez AM. Impact of Adverse Childhood Experiences on Intimate Partner Violence Perpetration among Sri Lankan Men. *PLOS ONE* 2015; **10**: e0136321.

12 Fulu E, Jewkes R, Roselli T, Garcia-Moreno C, UN Multi-country Cross-sectional Study on Men and Violence research team. Prevalence of and factors associated with male perpetration of intimate partner violence: findings from the UN Multi-country Cross-sectional Study on Men and Violence in Asia and the Pacific. *Lancet Glob Health* 2013; **1**: e187-207.

13 Narayan AJ, Labella MH, Englund MM, Carlson EA, Egeland B. The legacy of early childhood violence exposure to adulthood intimate partner violence: Variable- and person-oriented evidence. *J Fam Psychol JFP J Div Fam Psychol Am Psychol Assoc Div 43* 2017; **31**: 833–43.

14 Abrahams N, Jewkes R. Effects of South African men's having witnessed abuse of their mothers during childhood on their levels of violence in adulthood. *Am J Public Health* 2005; **95**: 1811–6.

15 Jewkes R, Dunkle K, Koss MP, *et al*. Rape perpetration by young, rural South African men: Prevalence, patterns and risk factors. *Soc Sci Med 1982* 2006; **63**: 2949–61.

16 Violence against children | Department of Economic and Social Affairs. https://sdgs.un.org/topics/violence-against-children (accessed June 8, 2025).

17 Laurin J, Wallace C, Draca J, Aterman S, Tonmyr L. Youth self-report of child maltreatment in representative surveys: a systematic review. *Health Promot Chronic Dis Prev Can Res Policy Pract* 2018; **38**: 37–54.

18 Mathews B, Pacella R, Dunne MP, Simunovic M, Marston C. Improving measurement of child abuse and neglect: A systematic review and analysis of national prevalence studies. *PLOS ONE* 2020; **15**: e0227884.

19 World Health Organization. Preventing child maltreatment : a guide to taking action and generating evidence / World Health Organization and International Society for Prevention of Child Abuse and Neglect. *Guide Sur Prév Maltraitance Enfants Interv Produire Données* 2006. https://apps.who.int/iris/handle/10665/43499 (accessed April 21, 2023).

20 UNICEF. International Classification of Violence against Children (ICVAC). UNICEF DATA. 2023; published online June 30. https://data.unicef.org/resources/international-classification-of-violence-against-children/ (accessed Sept 13, 2023).

21 UNICEF. Measuring Violence Against Children - Inventory and assessment of quantitative studies. UNICEF DATA. 2014; published online Nov 3. https://data.unicef.org/resources/measuring-violence-against-children-inventory-and-assessment-of-quantitative-studies-publication/ (accessed Jan 1, 2025).

22 Brauer M, Roth GA, Aravkin AY, *et al*. Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet* 2024; **403**: 2162–203.

23 Arora A. When Numbers Demand Action: Confronting the global scale of sexual violence against children. UNICEF DATA. 2024; published online Oct 9. https://data.unicef.org/resources/when-numbers-demand-action/ (accessed March 5, 2025).

24 Straus MA, Hamby SL, Finkelhor D, Moore DW, Runyan D. Identification of Child Maltreatment With the Parent-Child Conflict Tactics Scales: Development and Psychometric Data for a National Sample of American Parents. *Child Abuse Negl* 1998; **22**: 249–70.

25 Devries K, Knight L, Petzold M, *et al*. Who perpetrates violence against children? A systematic analysis of age-specific and sex-specific data. *BMJ Paediatr Open* 2018; **2**: e000180.

26 Chan KL. Are parents reliable in reporting child victimization? Comparison of parental and adolescent reports in a matched Chinese household sample. *Child Abuse Negl* 2015; **44**: 170–83.

27 Hardt J, Rutter M. Validity of adult retrospective reports of adverse childhood experiences: review of the evidence. *J Child Psychol Psychiatry* 2004; **45**: 260–73.

28 McPherson L, Gatwiri K, Graham A, *et al*. What Helps Children and Young People to Disclose their Experience of Sexual Abuse and What Gets in the Way? A Systematic Scoping Review. *Child Youth Care Forum* 2024; published online Sept 18. DOI:10.1007/s10566-024-09825-5.

29 Langeland W, Smit JH, Merckelbach H, de Vries G, Hoogendoorn AW, Draijer N. Inconsistent retrospective self-reports of childhood sexual abuse and their correlates in the general population. *Soc Psychiatry Psychiatr Epidemiol* 2015; **50**: 603–12.

30 Breton E, Kidman R, Behrman J, Mwera J, Kohler H-P. Longitudinal consistency of self-reports of adverse childhood experiences among adolescents in a low-income setting. *SSM - Popul Health* 2022; **19**: 101205.

31 UNICEF. Ethical Principles, Dilemmas and Risks in Collecting Data on Violence Against Children. UNICEF DATA. 2012; published online Oct 15. https://data.unicef.org/resources/ethical-dilemmas-risks-collecting-data-violence-children-findings-work-cp-merg-technical-working-group-violence-children/ (accessed Jan 30, 2025).

32 Bhatia A, Zinke-Allmang A, Bangirana CA, *et al.* Putting children's safety at the heart of violence research. *Nat Med* 2024; **30**: 2721–4.

33 Sofuoğlu Z, Sarıyer G, Ataman MG. CHILD MALTREATMENT IN TURKEY: COMPARISON OF PARENT AND CHILD REPORTS. *Cent Eur J Public Health* 2016; **24**: 217–22.

34 Sierau S, White LO, Klein AM, Manly JT, von Klitzing K, Herzberg PY. Assessing psychological and physical abuse from children's perspective: Factor structure and psychometric properties of the picture-based, modularized child-report version of the Parent-Child Conflict Tactics Scale - Revised (CTSPC-R). *PloS One* 2018; **13**: e0205401.

35 Hogan JN, Garcia AM, Tomko RL, Squeglia LM, Flanagan JC. Parent-Child Concordance and Discordance in Family Violence Reporting: A Descriptive Analysis from the Adolescent Brain Cognitive Development Study®. *J Interpers Violence* 2023; **38**: NP646–69.

36 Dunn EC, Ernst SC, Nishimi K, Choi KR. The Prevalence, Predictors, and Health Consequences of Disagreement in Reports of Child Maltreatment Exposure. *Child Psychiatry Hum Dev* 2024; published online May 30. DOI:10.1007/s10578-024-01721-2.

37 Boyd A, Golding J, Macleod J, *et al.* Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013; **42**: 111–27.

38 Fraser A, Macdonald-Wallis C, Tilling K, *et al.* Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013; **42**: 97–110.

39 Northstone K, Lewcock M, Groom A, *et al.* The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. *Wellcome Open Res* 2019; **4**: 51.

40 Golding, Pembrey, Jones, Team TAS. ALSPAC–The Avon Longitudinal Study of Parents and Children. *Paediatr Perinat Epidemiol* 2001; **15**: 74–87.

41 World Health Organization. World report on violence and health. 2002. https://www.who.int/publications/i/item/9241545615 (accessed June 8, 2025).

42 Shan G, Lou X, Wu SS. Continuity Corrected Wilson Interval for the Difference of Two Independent Proportions. *J Stat Theory Appl* 2023; **22**: 38–53.

43 Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol* 2009; **9**: 5.

44 Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005; **37**: 360–3.

45 Lee SJ, Lansford JE, Pettit GS, Bates JE, Dodge KA. Parental agreement of reporting parent to child aggression using the Conflict Tactics Scales. *Child Abuse Negl* 2012; **36**: 510–8.

46 Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. *Epidemiol Camb Mass* 2013; **24**: 1–9.

47 van Buuren, S. Flexible Imputation of Missing Data. Second Edition. Chapman and Hall 2018. https://doi.org/10.1201/9780429492259 (accessed March 2, 2025).

48 Houtepen LC, Heron J, Suderman MJ, Tilling K, Howe LD. Adverse childhood experiences in the children of the Avon Longitudinal Study of Parents and Children (ALSPAC). Wellcome Open Res. 2018 Aug 30;3:106. doi: 10.12688/wellcomeopenres.14716.1. PMID: 30569020; PMCID: PMC6281007.

49 Chan CS, Sarvet AL, Basu A, Koenen K, Keyes KM. Associations of intimate partner violence and financial adversity with familial homelessness in pregnant and postpartum women: A 7-year prospective study of the ALSPAC cohort. *PLOS ONE* 2021; **16**: e0245507.

50 Von Hippel PT. How to Impute Interactions, Squares, and Other Transformed Variables. *Sociol Methodol* 2009; **39**: 265–91.

51 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; **30**: 377–99.

52 Yakubovich AR, Heron J, Feder G, Fraser A, Humphreys DK. Intimate partner violence victimisation in early adulthood: psychometric properties of a new measure and gender differences in the Avon Longitudinal Study of Parents and Children. *BMJ Open* 2019; **9**: e025621.

53 Herbert A, Fraser A, Howe LD, *et al.* Categories of Intimate Partner Violence and Abuse Among Young Women and Men: Latent Class Analysis of Psychological, Physical, and Sexual Victimization and Perpetration in a UK Birth Cohort. *J Interpers Violence* 2023; **38**: 931–54.

54 Herbert A, Heron J, Barter C *et al.* Risk factors for intimate partner violence and abuse among adolescents and young adults: findings from a UK population-based cohort [version 3; peer review: 2 approved]. Wellcome Open Res. 2021, **5**:176 (https://doi.org/10.12688/wellcomeopenres.16106.3) (accessed March 8, 2023).

55 Nation A, Pacella R, Monks C, Mathews B, Meinck F. Prevalence of violence against children in the United Kingdom: A systematic review and meta-analysis. *Child Abuse Negl* 2023; **146**: 106518.

56 Nace A, Maternowska C, Fernandez B, Cravero K. The Violence Against Children Surveys (VACS): Using VACS data to drive programmes and policies. *Glob Public Health* 2022; **17**: 2807–25.

57 Turner HA, Finkelhor D, Ormrod R. The effect of lifetime victimization on the mental health of children and adolescents. *Soc Sci Med 1982* 2006; **62**: 13–27.

58 Guide_to_DHS_Statistics_DHS-7_v2.pdf. https://dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_DHS-7_v2.pdf (accessed June 8, 2025).

59 Dunne MP, Zolotor AJ, Runyan DK, *et al*. ISPCAN Child Abuse Screening Tools Retrospective version (ICAST-R): Delphi study and field testing in seven countries. *Child Abuse Negl* 2009; **33**: 815–25.

60 Hamby SL, Finkelhor D, Ormrod R, Turner H. THE JUVENILE VICTIMIZATION QUESTIONNAIRE (JVQ): ADMINISTRATION AND SCORING MANUAL. .

61 Barnes M, Szilassy E, Herbert A, *et al*. Being silenced, loneliness and being heard: understanding pathways to intimate partner violence & abuse in young adults. a mixed-methods study. *BMC Public Health* 2022; **22**: 1562.

62 Crick DCP, Halligan SL, Howe LD, *et al*. Associations between Adverse Childhood Experiences and the novel inflammatory marker glycoprotein acetyls in two generations of the Avon Longitudinal Study of Parents and Children birth cohort. *Brain Behav Immun* 2022; **100**: 112–20.

63 Goncalves Soares A, Zimmerman A, Zammit S, Karl A, Halligan SL, Fraser A. Abuse in Childhood and Cardiometabolic Health in Early Adulthood: Evidence From the Avon Longitudinal Study of Parents and Children. *J Am Heart Assoc* 2021; **10**: e021701.

64 Bauer A, Hammerton G, Fraser A, Fairchild G, Halligan SL. Associations between developmental timing of child abuse and conduct problem trajectories in a UK birth cohort. *BMC Psychiatry* 2021; **21**: 89.

Appendix

Appendix 1. VAC instruments and examples of behaviors per forms of abuse

Instrument, first year	Type of abuse	Respondent	Examples of behaviors
ACE,1998[3]	Physical, psychological, sexual	Self	Did a parent or other adult in the household often or very often push, grab, shove or slap you; Swear at, insult, or put you down; Any adult or child 5 years older touch or fondle you in a sexual way
CTSPS,1998[24]	Physical, psychological, sexual	Parent	Did parent hit on the bottom with something like a belt, hairbrush, a stick or some other hard object; Swore or cursed at; Did any adult or older child forced to have sex
MICS, 1995[21]	Physical, psychological	Parent	Did parent hit or slapped on the face, head or ears; Called him/her dumb, lazy or another name like that
VACS, 2012[56]	Physical, psychological, sexual	Self	Parent or adult caregiver hit you on the bottom or elsewhere with a bare hand or a hard object; Threatened to get rid of you; Anyone touched in a sexual way without your permission (includes fondling, pinching, grabbing, or touching you on or around your sexual body parts), but did not try and force you to have sex
LTVH, 2005[57]	Physical, sexual	Self	Parent/caregiver/adult choked, smothered, tried or attempted to drown, or burned intentionally; Anyone -male or female- ever forced or coerced you to engage in unwanted sexual activity (By unwanted sexual activity, we mean vaginal, oral, or anal intercourse, or has anyone inserted an object or their fingers in your anus or vagina)
DHS, 2006[23,58]	Physical, psychological, sexual	Parent (physical and psychological), self (sexual)	Parent or anyone else in the household beat up with an implement over and over as hard as one could; Shouted, yelled at, or screamed at; Anyone forced or unwanted sex
ISPCAN, 2009[59]	Physical, psychological, sexual	Parent/self	Parent/caregiver put hot pepper, soap, or spicy food in mouth (to cause pain); Use public humiliating to discipline him or her; Adult forced to have sexual intercourse
JVQ, 2004[60]	Physical, psychological, sexual	Self	Grown-up in your life hit, beat, kick, or physically hurt you in any way; Got scared or feel really bad because called you names; Grown-up you know touch your private parts when they shouldn't have or make you touch their private parts

Appendix 2. Young adult's experience of child physical abuse in ALSPAC surveys

Question	Respondents	Age of the child	Who used it
Since the last time we interviewed you: Your partner was physically cruel to your children	Mother, partner	18 weeks gestation, 8 weeks, 8 months, 21 months, 33 months, 47 months, 5 years, 6 years, 9 years, 11 years, 18 years, 19 years	Barnes et al. 2022[61], Houtepen et al. 2018, Crick et al. 2023[62], Herbert et al. 2023
Since the last time we interviewed you: You were physically cruel to your children?	Mother, partner	8 months, 21 months, 33 months, 47 months, 5 years, 6 years, 9 years, 11 years, 18 years, 19 years	Barnes et al. 2022, Houtepen et al. 2018, Crick et al. 2023, Herbert et al. 2023
Before the age of 11, how often did an adult in your family (anyone you consider to be a family member): c Push, grab or shove you; d Smack you for discipline; g Actually kick, punch, or hit you with something that could hurt you or physically attack you in another way; h Hit you so hard it left you with bruises or marks	Child	22 years	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Goncalves et al. 2021[63], Bauer et al. 2021[64], Herbert et al. 2023
Between the ages of 11 and 17, how often did an adult in your family (anyone you consider to be a family member): c Push, grab or shove you; d Smack you for discipline; g Actually kick, punch, or hit you with something that could hurt you or physically attack you in another way; h Hit you so hard it left you with bruises or marks	Child	22 years	Houtepen et al. 2018, Barnes et al. 2022, Crick et al. 2023
When I was growing up: b People in my family hit me so hard that it left me with bruises or marks	Child	23 years	Herbert et al. 2023
Before you were 19, a parent was physically cruel to you:	Child	27 years	
Since x period of time child was physically hurt by someone?	Mother	18 months, 30 months, 42 months, 57 months, 5 years, 6 years, 8 years	
When she has temper tantrums how often do you: vi slap or hit her	Mother	57 months	

Question	Respondents	Age of the child	Who used it
How often do you slap or hit child?	Mother	9 years 7 months	

Appendix 3. Young adult's experience of child psychological abuse in ALSPAC surveys

Question	Age of child	Respondents	Who used it
Since the last time we interviewed you: Your partner was emotionally cruel to your children?	18 weeks gestation, 8 weeks, 8 months, 21 months, 33 months, 47 months, 5 years, 6 years, 9 years, 11 years, 18 years, 19 years	Mother, partner	Barnes et al. 2022, Houtepen et al. 2018, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
Since the last time we interviewed you: You were emotionally cruel to your children?	8 months, 21 months, 33 months, 47 months, 5 years, 6 years, 9 years, 11 years, 18 years, 19 years	Mother, partner	Barnes et al. 2022, Houtepen et al. 2018, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
Before the age of 11, how often did an adult in your family (anyone you consider to be a family member): a Shout at you; b Say hurtful or insulting things to you; f Threaten to kick, punch, or hit you with something that could hurt you or physically attack you in another way	22 years	Child	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Goncalves et al. 2021, Bauer et al. 2021, Herbert et al. 2023
Between the ages of 11 and 17, how often did an adult in your family (anyone you consider to be a family member): a Shout at you; b Say hurtful or insulting things to you; e Punish you in a way that seemed cruel; f Threaten to kick, punch, or hit you with something that could hurt you or physically attack you in another way	22 years	Child	Houtepen et al. 2018, Barnes et al. 2022, Crick et al. 2023, Bauer et al. 2021, Herbert et al. 2023
When I was growing up: c I felt that someone in my family hated me	23 years	Child	
Before you were 19, a parent was emotionally cruel to you:	27 years	Child	

Appendix 4. Young adult's experience of child sexual abuse in ALSPAC surveys

Question	Age of child	Respondent	Who used it
Before the age of 11, were you touched in a sexual way by an adult or an older child or were you forced to touch an adult or older child in a sexual way when you did not want to?	22 years	Child	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Goncalves et

Question	Age of child	Respondent	Who used it
			al. 2021, Bauer et al. 2021, Herbert et al. 2023
Before the age of 11, did an adult or an older child force you or attempt to force you into any sexual activity by threatening you or holding you down or hurting you in some way when you did not want to?	22 years	Child	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Goncalves et al. 2021, Bauer et al. 2021, Herbert et al. 2023
Between the ages of 11 and 17, were you touched in a sexual way by an adult or an older child or were you forced to touch an adult or older child in a sexual way when you did not want to?	22 years	Child	Houtepen et al. 2018, Barnes et al. 2022, Crick et al. 2023, Bauer et al. 2021, Herbert et al. 2023
Between the ages of 11 and 17, did an adult or an older child force you or attempt to force you into any sexual activity by threatening you or holding you down or hurting you in some way when you did not want to?	22 years	Child	Houtepen et al. 2018, Barnes et al. 2022, Crick et al. 2023, Bauer et al. 2021, Herbert et al. 2023
When I was growing up: d Someone molested me (sexually)	23 years	Child	Herbert et al. 2023
Before you were 19 you were sexually abused	27 years	Child	
Since x period of time child was sexually abused/assaulted	18 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
	30 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
	42 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al.

Question	Age of child	Respondent	Who used it
			2023, Herbert et al. 2023
	57 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
	5 years 9 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
	6 years 9 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023
	8 years 7 months	Mother	Houtepen et al. 2018, Barnes et al. 2022, Herbert et al. 2021, Crick et al. 2023, Herbert et al. 2023

Appendix 5. Definition of mother's experience of child abuse and neglect

Type of abuse	Question	Definition	Age of child at round
Physical	Have any of these ever happened to you? L. Your parents hurt you	Abused if mother replied yes	12 weeks of gestation
	Were you physically abused (e.g. beaten) as a child? If yes, who abused you? Mother, father		2 years and 9 months
Psychological	Before you were 17: A parent was emotionally cruel to you?	Abused if mother replied yes	32 weeks of gestation
Sexual	Did anyone ever ..... before you were 16? touch or fondle your body, including your breast or genitals, or attempt to arouse you sexually; try to have you arouse them, or touch their body in a sexual way; rub their genitals against your body in a sexual way; have sexual intercourse with you; try to put their penis into your mouth	Abused if experienced any of the acts before she was 16 years old, and she did not want any of these experiences to happen no matter the perpetrator, or the perpetrator was anyone aside from her boyfriend or girlfriend no matter whether she wanted it to happen	
	Who was involved? Boyfriend, girlfriend, parent, sibling, family friend, other relative, other person, stranger		
	Did you wanted this to happen?		32 weeks of gestation
Neglect	Did you feel neglected emotionally during your childhood?	Neglected if mother replied yes to either question	
	Were you physically neglected as a child (e.g. not fed or clothed properly)?		2 years and 9 months
Witnessed parental IPV	How would you describe the relationship between your mother and father when you were growing up? I. Violent	Witnessed if mother replied yes	2 years and 9 months

Appendix 6. Comparison of complete cases and MICE imputation

			Complete cases		Imputation	
Variable	**N**	**Percent missing**	**Mean/Proportion**	**Standard error**	**Mean/Proportion**	**Standard error**
Female child	3951	0.00	0.64	0.01	0.64	0.01
Child report of CPA at 27	3951	0.00	0.06	0.00	0.06	0.00
Child report of psychological abuse at 27	3951	0.00	0.12	0.01	0.12	0.01
Child report of sexual abuse at 27	3942	0.23	0.02	0.00	0.02	0.00
Mom report of CPA	3951	0.00	0.05	0.00	0.05	0.00
Mom report of psychological abuse	3951	0.00	0.13	0.01	0.13	0.01
Mom report of physical IPV cum.	3951	0.00	0.09	0.00	0.09	0.00
Mom report of psychological IPV cum.	3951	0.00	0.28	0.01	0.28	0.01
Mothers age at pregnancy	3900	1.29	28.92	0.07	28.92	0.07
N other children at pregnancy	3831	3.04	0.74	0.01	0.74	0.01
Married mom at pregnancy	3901	1.27	0.84	0.01	0.83	0.01
White child	3804	3.72	0.97	0.00	0.96	0.00
Postpartum depression score 18 weeks gest.	3623	8.3	6.23	0.07	6.27	0.07
Postpartum depression score 32 weeks gest.	3764	4.73	6.47	0.08	6.54	0.08
Postpartum depression score 8 weeks	3774	4.48	5.67	0.07	5.71	0.07
Partner's postpartum depression 18 weeks gest.	3189	19.29	3.92	0.07	4.01	0.06
Child's birth weight in grams	3901	1.27	3413.13	8.59	3412.84	8.53
Gestational age in weeks	3951	0.00	39.50	0.03	39.50	0.03
Mother homeless at pregnancy	3722	5.8	0.02	0.00	0.02	0.00
N alcoholic beverages 32 weeks gest	2293	41.96	1.68	0.07	1.83	0.06
N alcoholic beverages 3 years and 11m	3644	7.77	4.86	0.10	4.83	0.10
Mom's psychological child abuse	3851	2.53	0.07	0.00	0.07	0.00
Mom's physical child abuse	3926	0.63	0.04	0.00	0.04	0.00
Mom's neglect	3627	8.2	0.20	0.01	0.20	0.01
Mom's witnessing parental IPV	3463	12.35	0.11	0.01	0.11	0.00
Mom's sexual child abuse	3580	9.39	0.25	0.01	0.25	0.01
Child reported abuse	3951	0.00	0.13	0.01	0.13	0.01

			Complete cases		Imputation	
Variable	N	Percent missing	Mean/Proportion	Standard error	Mean/Proportion	Standard error
Mom reported any abuse	3951	0.00	0.15	0.01	0.15	0.01
Mom education: High school	3863	2.23	0.35	0.01	0.35	0.01
Mom education: More than high school	3863	2.23	0.47	0.01	0.47	0.01
Mom education: No high school	3863	2.23	0.18	0.01	0.18	0.01
Partner education: High school	3798	3.87	0.21	0.01	0.21	0.01
Partner education: More than high school	3798	3.87	0.55	0.01	0.54	0.01
Partner education: No high school	3798	3.87	0.24	0.01	0.25	0.01
Mom social class: I & II	3384	14.35	0.45	0.01	0.43	0.01
Mom social class: III	3384	14.35	0.47	0.01	0.48	0.01
Mom social class: IV & V	3384	14.35	0.08	0.00	0.09	0.00
Partner social class: I & II	3587	9.21	0.53	0.01	0.52	0.01
Partner social class: III	3587	9.21	0.38	0.01	0.39	0.01
Partner social class: IV & V	3587	9.21	0.09	0.00	0.10	0.00
Financial difficulty 32 weeks gest: 1-2	3774	4.48	0.26	0.01	0.25	0.01
Financial difficulty 32 weeks gest: 3-7	3774	4.48	0.23	0.01	0.23	0.01
Financial difficulty 32 weeks gest: 8 or more	3774	4.48	0.08	0.00	0.08	0.00
Financial difficulty 32 weeks gest: None	3774	4.48	0.43	0.01	0.43	0.01
Financial difficulty 8 months: 1-2	3791	4.05	0.26	0.01	0.26	0.01
Financial difficulty 8 months: 3-7	3791	4.05	0.27	0.01	0.27	0.01
Financial difficulty 8 months: 8 or more	3791	4.05	0.11	0.01	0.11	0.01
Financial difficulty 8 months: None	3791	4.05	0.36	0.01	0.36	0.01
Financial difficulty 1 year 6m: 1-2	3680	6.86	0.27	0.01	0.27	0.01
Financial difficulty 1 year 6m: 3-7	3680	6.86	0.25	0.01	0.25	0.01
Financial difficulty 1 year 6m: 8 or more	3680	6.86	0.11	0.01	0.11	0.00
Financial difficulty 1 year 6m: None	3680	6.86	0.37	0.01	0.37	0.01
Financial difficulty 2 years 9m: 1-2	3626	8.23	0.25	0.01	0.25	0.01
Financial difficulty 2 years 9m: 3-7	3626	8.23	0.25	0.01	0.25	0.01
Financial difficulty 2 years 9m: 8 or more	3626	8.23	0.11	0.01	0.11	0.01
Financial difficulty 2 years 9m: None	3626	8.23	0.39	0.01	0.39	0.01

			Complete cases		Imputation	
Variable	N	Percent missing	Mean/Proportion	Standard error	Mean/Proportion	Standard error
Difficulty affording food 32 weeks gest: Fairly difficult	3774	4.48	0.04	0.00	0.04	0.00
Difficulty affording food 32 weeks gest: No difficulty	3774	4.48	0.82	0.01	0.82	0.01
Difficulty affording food 32 weeks gest: Some difficulty	3774	4.48	0.13	0.01	0.13	0.01
Difficulty affording food 32 weeks gest: Very difficult	3774	4.48	0.01	0.00	0.01	0.00
Difficulty affording heating 32 weeks gest: Fairly difficult	3774	4.48	0.05	0.00	0.06	0.00
Difficulty affording heating 32 weeks gest: No difficulty	3774	4.48	0.78	0.01	0.77	0.01
Difficulty affording heating 32 weeks gest: Some difficulty	3774	4.48	0.15	0.01	0.15	0.01
Difficulty affording heating 32 weeks gest: Very difficult	3774	4.48	0.02	0.00	0.02	0.00

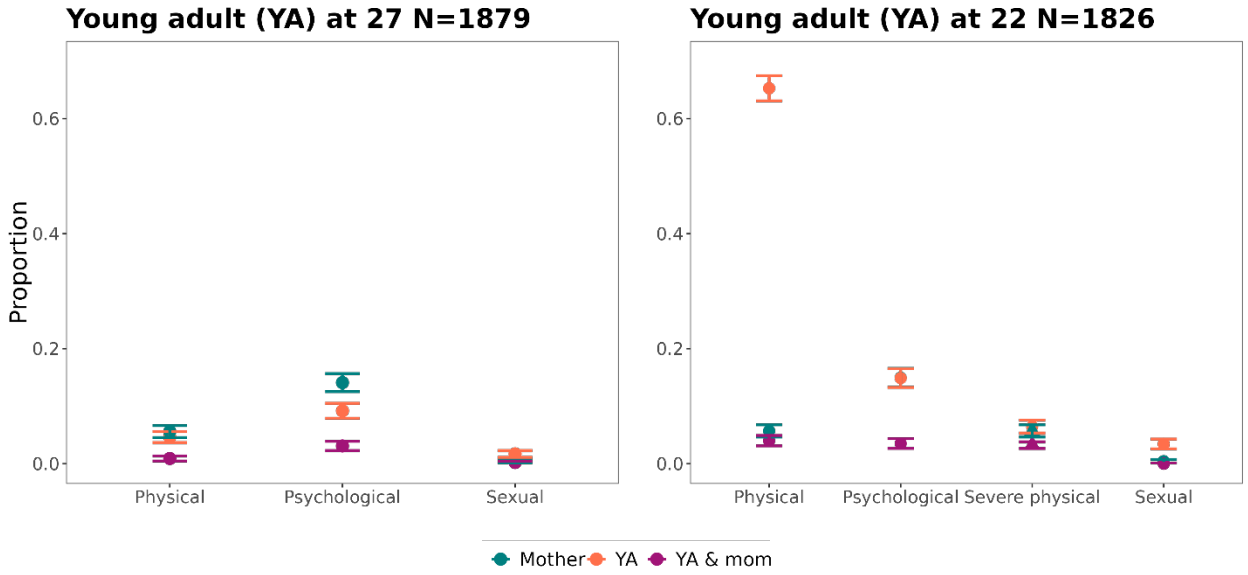
Appendix 7. Multivariate logistic regression on mother reporting either physical or psychological abuse in imputed and complete cases samples

	Imputed, any		Complete, any	
	OR (95% CI)	**P value**	**OR (95% CI)**	**P value**
Child predictors				
Female child	0.69 (0.51-0.93)	0.01	0.77 (0.53-1.11)	0.17
White child	1.23 (0.65-2.35)	0.52	1.67 (0.70-3.89)	0.24
Mother predictors at pregnancy				
Age	1.03 (0.99-1.07)	0.10	1.01 (0.97-1.06)	0.61
N other children	1.26 (1.06-1.51)	0.01	1.29 (1.01-1.66)	0.05
Education				
High school	0.75 (0.52-1.07)	0.11	0.56 (0.37-0.86)	0.01
No high school	0.76 (0.50-1.16)	0.20	0.70 (0.41-1.20)	0.19
Married	1.19 (0.85-1.67)	0.31	1.01 (0.65-1.57)	0.95
Social class				
III	0.84 (0.60-1.19)	0.34	0.83 (0.55-1.25)	0.37
IV & V	1.00 (0.57-1.77)	1.00	0.92 (0.48-1.81)	0.82
Psychological child abuse	0.83 (0.50-1.36)	0.46	0.58 (0.31-1.06)	0.08
Physical child abuse	2.16 (1.12-4.16)	0.02	2.94 (1.32-7.15)	0.01
Neglect	1.00 (0.69-1.44)	1.00	1.01 (0.64-1.60)	0.96
Witnessed parental IPV	1.59 (1.02-2.48)	0.04	1.36 (0.80-2.35)	0.27
Sexual child abuse	1.34 (0.99-1.82)	0.06	1.51 (1.04-2.22)	0.03
N	945		623	
AIC	1226.19		805.19	

Appendix 8. Multivariate logistic regression on mother's reporting abuse when young adult reported either physical or psychological abuse in any of the rounds "Life at 22+" or "Life at 27+"

	Life at 27+, imputed		Life at 22+ and 27+, imputed	
	OR (95% CI)	**P value**	**OR (95% CI)**	**P value**
Child predictors				
Female child	0.69 (0.51-0.93)	0.01	0.90 (0.67-1.21)	0.49
White child	1.23 (0.65-2.35)	0.52	1.57 (0.76-3.24)	0.22
Mother predictors at pregnancy				
Age	1.03 (0.99-1.07)	0.10	1.04 (1.01-1.08)	0.02
N other children	1.26 (1.06-1.51)	0.01	1.19 (1.00-1.41)	0.04
Education				
High school	0.75 (0.52-1.07)	0.11	0.71 (0.50-1.01)	0.06
No high school	0.76 (0.50-1.16)	0.20	0.61 (0.40-0.93)	0.02
Married	1.19 (0.85-1.67)	0.31	0.95 (0.66-1.36)	0.77
Social class				
III	0.84 (0.60-1.19)	0.34	0.98 (0.69-1.39)	0.92
IV & V	1.00 (0.57-1.77)	1.00	1.15 (0.66-1.99)	0.63
Psychological child abuse	0.83 (0.50-1.36)	0.46	1.13 (0.69-1.85)	0.63
Physical child abuse	2.16 (1.12-4.16)	0.02	1.43 (0.77-2.67)	0.26
Neglect	1.00 (0.69-1.44)	1.00	1.37 (0.95-1.97)	0.09
Witnessed parental IPV	1.59 (1.02-2.48)	0.04	1.06 (0.67-1.67)	0.81
Sexual child abuse	1.34 (0.99-1.82)	0.06	1.48 (1.10-2.01)	0.01
N	945		932	
AIC	1226.19		1251.30	

Appendix 9. Proportion of young adults experiencing abuse by respondent and survey round when sample of mothers are those who replied to all survey rounds



Appendix 10. Proportions of physical, psychological and sexual abuse per respondent when sample of mothers are those who replied to all survey rounds

		Young Adult report		Mother Report		Young Adult vs. mother		
Sample	Form of abuse	M	Proportion (95% CI)	M	Proportion (95% CI)	Cohen's Kappa (95% CI)	PABAK (95% CI)	Fisher (p-value)
At 27	Physical	0	0.05 (0.04-0.06)	0	0.06 (0.05-0.07)	0.13 (0-0.26)	0.83 (0.8-0.86)	4.7 (0)
At 27	Psychological	0	0.09 (0.08-0.11)	0	0.14 (0.13-0.16)	0.17 (0.09-0.25)	0.66 (0.62-0.69)	3.65 (0)
At 27	Sexual	3	0.02 (0.01-0.02)	0	0 (0-0.01)	0.09 (-0.2-0.38)	0.96 (0.95-0.97)	17.38 (0.01)
At 22	Physical	0	0.65 (0.63-0.67)	0	0.06 (0.05-0.07)	0.01 (-0.03-0.04)	-0.26 (-0.3--0.21)	1.27 (0.29)
At 22	Severe physical	0	0.06 (0.05-0.08)	0	0.06 (0.05-0.07)	0.1 (-0.02-0.22)	0.79 (0.76-0.82)	3.13 (0)
At 22	Sexual	20	0.03 (0.03-0.04)	0	0 (0-0.01)	-0.01 (-0.24-0.23)	0.92 (0.9-0.94)	0 (1)
At 22	Psychological	0	0.15 (0.13-0.17)	0	0.15 (0.13-0.17)	0.1 (0.02-0.18)	0.54 (0.5-0.58)	1.98 (0)

Note: M= Missing. In sample "At 22", mothers; report of physical violence is the same for physical and severe physical, it is being compared against two different definitions of physical abuse from the child questionnaire.

Appendix 11. Multivariate logistic regression on mother's reporting abuse with two imputed pairs of young adults and mothers

	Mothers with at least 2 rounds, any		Mothers with all rounds, any	
	OR (95% CI)	P value	OR (95% CI)	P value
Child predictors				
Female child	0.69 (0.51-0.93)	0.01	0.57 (0.36-0.92)	0.02
White child	1.23 (0.65-2.35)	0.52	1.19 (0.35-4.12)	0.78
Mother predictors at pregnancy				
Age	1.03 (0.99-1.07)	0.10	1.01 (0.96-1.07)	0.61
N other children	1.26 (1.06-1.51)	0.01	1.37 (1.02-1.85)	0.04
Education				
High school	0.75 (0.52-1.07)	0.11	0.78 (0.44-1.38)	0.39
No high school	0.76 (0.50-1.16)	0.20	0.75 (0.36-1.56)	0.44
Married	1.19 (0.85-1.67)	0.31	1.22 (0.67-2.22)	0.52
Social class				
III	0.84 (0.60-1.19)	0.34	0.79 (0.45-1.39)	0.42
IV & V	1.00 (0.57-1.77)	1.00	1.36 (0.45-4.10)	0.59
Psychological child abuse	0.83 (0.50-1.36)	0.46	0.92 (0.40-2.08)	0.84
Physical child abuse	2.16 (1.12-4.16)	0.02	4.16 (1.12-15.45)	0.03
Neglect	1.00 (0.69-1.44)	1.00	1.21 (0.65-2.24)	0.55
Witnessed parental IPV	1.59 (1.02-2.48)	0.04	0.90 (0.42-1.95)	0.80
Sexual child abuse	1.34 (0.99-1.82)	0.06	1.11 (0.69-1.80)	0.67
N	945		425	
AIC	1226.19		510.75	

Note: The first regression is on the sample of mothers that respondent to at least 2 rounds of surveys. The second regression is on the sample of mothers that respondents to all rounds of survey.

Chapter 2: Differential health loss valuation by sex of the respondent in the GBD study

Introduction

Overview

In this chapter we analyze whether sex of the respondent influences health loss as measured through disability weights (DW) in the Global Burden of Disease (GBD) study. For that purpose, we first describe what disability weights are used for, their evolution in the 30 years since they were first released, and the methodology to estimate them. Then, we summarize past literature that analyzed whether sex of the respondents or other characteristics affect the estimation of disability weights to highlight the research gap addressed by this analysis. The rest of the chapter is dedicated to estimating sex disaggregated disability weights and understanding the differences in health loss valuation by sex.

Disability adjusted life years and evolution of disability weights

When working towards the improvement of population health, policy makers need to decide how to distribute finite resources among different health issues through the implementation of health policies, health systems strengthening and intervention programs[1]. Summary measures of population health can be used to help inform these decisions, as they are able to track improvements in population health over time and in comparison to other populations. These are single health metrics that combine age-specific death and functional health loss. They have been classified in two families: health expectancies and health gaps. Health expectancies metrics such as health-adjusted life expectancy (HALE), and quality-adjusted life year (QALY) measure the overall, average level of health in a population, while health gap metrics such as disability-adjusted life years (DALY) measure years of life lost due to death and deteriorated health[2–4].

DALYs were first presented in the World Bank's World development report of 1993 as a metric to both measure the burden of disease, and be used in cost-effectiveness analysis[5]. After its release,-DALYs have been recommended for health economic evaluations in low-and middle-income countries by the World Bank and the Gates Foundation[6,7]. Today, DALYs represent a comprehensive summary

metric of population health across age groups, sexes and locations. In one metric, DALYs assess the burden of morbidity, measured through Year of Life with Disability (YLD), and mortality, measured through Years of Life Lost (YLL), of 369 diseases and injuries across 204 countries[8]. To estimate YLD, disability weights need to be estimated for all mutually exclusive consequences of disease and injuries -also known as sequelae- and then multiplied by the prevalence of each sequela per country year[8]. In other words, a unique disability weight per sequela is used across country, time and population demographic characteristics.

Disability weights for the GBD study are continuous values between 0 and 1 where 0 represents perfect health and 1 represents death. They have gone through three major updates since they were first presented in 1994: in 1996, 2010, and 2013[9–11]. To understand their evolution over time, we use Haagsma et al.'s 2014 conceptual model of assessing disability weights and design choices[12]. Four major design choices are described by Haagsma et al. in assessing heath loss valuation. First, when defining the health states to which disability weights would be assigned, choices must be made on whether these would be disease-specific or described in generic terms. Disease-specific conditions include the label of the disease, and they describe the health effects and causes of the condition. Generic health state descriptions focus on functional health loss independent of the disease that could be causing it, they include multiple dimensions of loss such as mobility, physical activity, social functioning, etc. The second choice must be made on whether to present the health state as temporal or as a chronic issue. The third choice is who will be making the health loss valuations, whether experts in the field, people with first experience of a certain disease or the public. And finally, there is the choice of which valuation method to use. There are many different methods used to derive these health valuations. Briefly, they can be categorized as rating scales methods and trade-off methods following Essink-Bot's and Bonsel's framework[13]. In rating scale methods, respondents are asked to rank different health states in order of worst imaginable health state to best, the most common example is the visual analogue scaling method (VAS). This method is easy to use and understand. However, VAS disability weights from large sample surveys are relatively high for relative mild conditions, and they do not have interval properties[13,14]. In trade-off methods respondents are asked to hypothetically decide between "something valuable" in exchange for preventing health loss, they include standard gamble, time-trade off (TTO), and personal trade-off (PTO) methods. One of the PTO questions used in the GBD study of 1996, for example, asked individuals to choose between a program that would extend the life of a population of healthy individuals, versus a program that

extended the life of individuals with a certain health condition[11]. These methods rely on "abstract and cognitively demanding thought experiments"[15]. Meaning choices between health states could be reflecting noise from respondents not answering the questions based on their actual preferences but on choosing randomly because they do not understand the exercise.

Initially, in the first GBD study of 1990, a small group of 20 health experts were invited to assign weights to a list of six different disability classes based on limitations to daily living, instrumental daily living activities, procreation, occupation, education and recreation. The least severe disability class, for example, was defined as "limited ability to perform at least one activity in one of the following areas: recreation, education, procreation or occupation"[11]. Then, disability weights for a sequela were estimated by assigning distribution of incident cases across the different disability classes[11]. In 1996, the estimation was revised to incorporate updates in the methodology. Murray et al. developed a protocol to valuate 22 indicator conditions that would encompass a wide range of disability severities and health states. The main valuation method was person-trade-off, two different types of PTO questions were asked to assure consistency between individual responses. The results from the person-trade-off exercise were averaged across participants, and defined the disability weights for the 22 conditions that were then divided into seven disability classes. This exercise was conducted by a small number of experts, were panel members with the lowest and highest valuations were asked to provide motivations for their choices, followed by a group discussion, and an opportunity to revise their original choices. In the second part of the protocol, participants were asked to map 483 sequelae across the seven classes reflecting either proportion of time spent or percentage of cases in different disability classes[16,17].

Then in 2010, the estimation of disability weights was adapted to follow developments in the field as well as the commentary and debate that surged from the widespread use of the GBD study.
Salomon et al. developed a new disability weights methodology in which they decided on a number of choices[9]. First, the health states would be disease-specific and not include label of the disease to avoid preconceived notions of a disease and focus only on the body and mental functioning failures of the health states. This contrasted with the 1996 methodology, where experts knew which diseases they were being asked about in the PTO exercises. For the GDB DW study of 2010, 220 health states and lay descriptions were developed that were mapped to an expanded list of 291 diseases and injuries and 1,160 sequelae, many more than the 483 sequelae of 1996. Second, both acute or

chronic temporality were presented for the same health state. Third, given that DALYs are used to increase health of populations, Salomon et al. concluded that DW estimates should be derived from population surveys instead of health experts assigning values to health loss for disability. For the GBD DW study of 2010, household, phone and web population surveys were rolled out to elicit health state preferences. In all, disability weights were derived from the responses of 30,000 survey participants, a meaningful contrast to the 100 health experts from the 1996 study. Fourth, because DW were estimated from the general population, the methods used to valuate health were purposely designed to be easy for people of all educational backgrounds to understand[9]. As such, Salomon et al. developed the paired comparison method -which falls under the rating scale family of methods- as the primary tool to elicit health state preferences from population, and developed a new trade-off method known as population health equivalence questions to anchor disability weights values from 0 to 1.

Finally, in 2013, new survey data from a European based disability weight study were pooled with the GBD DW measurement study in 2010 to estimate updated disability weights[10]. As of 2025, the GBD study still uses disability weights derived in 2013.

GBD disability weights methodology

Table 1 presents the characteristics of the GBD 2010 DW study and the European DW study. In the GBD 2010 DW study, data was collected from household surveys in Bangladesh, Indonesia, Peru, Tanzania; telephone surveys in the United States; and open access web-based surveys in 167 countries around the world. Household surveys were multistage, randomly selected stratified samples with probabilities of being selected proportional to population size. In the United States researchers conducted interviews over the phone using the Behavioral Risk Factor Surveillance System as contacts. The web survey did not follow a random sample design, instead respondents were recruited through news and editorials in scientific journals, using the investigators' professional networks and other similar form of convenience sampling. Further detail on the sample and data collection can be found in Salomon et al. 2010[9].

Table 1. Study characteristics used to derive GBD 2013 disability weights

Study	Years	Geographical Scope	Age of Respondents	Number of Respondents	Health states
GBD 2010	2009-2010	Bangladesh, Indonesia, Peru, Tanzania, and United States	18 years and older	13,902	284
	2010-2011	167 countries	18 years and older	16,328	
European	2013	Hungary, Italy, Netherlands, and Sweden	18 to 65 years	30,660	255

In 2013, the European Centre for Disease Prevention and Control ran the European DW study with the intention of estimating DW from a European representative sample. Researchers selected four countries that represented Eastern, Southern, Central and Northern Europe and respondents were randomly selected from internet panel lists to match certain demographic characteristics of the population of 18 to 65 years old in each country, see Haagsma et al. 2015 for details[18].

The GBD disability weights methodology can be summarized in six concrete steps. First, health states that would be valued were identified and lay health state descriptions developed for them. Then health state preferences were estimated with paired comparison (PC) data. In step three population health equivalence (PHE) questions were used to derive disability weights between 0 and 1 for a selected smaller set of conditions. Fourth, the disability weights estimated in step 3 were used to anchor health preferences elicited in step two as values between 0 and 1. In step five, global disability weights are adjusted based on variation across countries. Finally, uncertainty estimates were derived from 1,000 bootstrap samples[9,10,19,20]. The following sections describe each of these steps in detail.

Step 1: Health states and health state lay descriptions

Both paired comparison and population health equivalence questions used health states lay descriptions to assert people's preferences on health states. Lay descriptions were less than 35 words in the GBD 2010 DW study, and less than 70 words in the European DW study. Descriptions were phrased in simple, non-clinical vocabulary, and were intended to emphasize major functional consequences and symptoms associated with each health state. In the GBD, every sequela is associated with a health state, therefore these descriptions needed to be general enough to be used

by themselves or in combined for each sequela -the GBD 2021 study used more than 3,000 sequelae[21].

In 2010, there were 1,160 sequelae in the GBD study. To select the final 220 health states associated with all of these, researchers took a disease-by-disease approach. First, they identified sequela per disease, then the sources of the sequela and their commonalities, while iteratively consulting with experts. Experts were provided with standardized worksheets that included dimensions of functional health and symptoms that could be used to describe health states. These descriptions were then reviewed and edited to align across each other, keeping main functional health loss characteristics and being as parsimonious as possible[9].

For the European DW study, 30 health states were modified and 17 were added. Additions were incorporated as a result of mapping health outcome trees for communicable disease and cross checking whether a health state existed for each severity class in the trees[18]. As for the modifications, Salomon et al. 2013 and colleagues gathered feedback from different experts following the publication of the GBD 2010 DW study and prior to the data collection of the European study. Lay descriptions were modified if they were missing key components or if descriptions differed too much to other similar conditions[10].

Step 2: Paired comparison questions and analysis

PC questions were the primary data used to elicit preferences of health states. They contained two hypothetical scenarios: in one a person has a health state A, and in the other a person suffers from health state B. Respondents were asked to choose which of the two individuals was the healthiest. The health state condition associated to each person was selected from a pool of 284 health states in the GBD 2010 DW study, and a pool of 255 health states in the European study. The health states being compared would randomly change per respondent. An example of the PC questions can be found in Salomon et al. 2010 appendix[9,18].

PC were analyzed using probit regression, with the following model:

$$choosefirst = \beta_0 + \beta_1 picked1 + \beta_2 picked2 + \cdots + \beta_{366} picked366$$

Where:

- choosefirst is either 0 or 1, if 1 then the first health state presented in the PC question was chosen as healthier.

- picked`i` is either -1, 0 or 1. i denotes the ID of the health state, if value is -1 then this health state was the second in the PC question. If value is 0 then this health state was not in the PC question. If value is 1, then his health state was the first one in the PC.

In the GBD 2010 DW study, five versions of the surveys were rolled out, in one version respondents were asked five PC questions, while in all other versions, respondents were asked 15 PC questions. Versions differed given the health states presented or the temporality of the health conditions. In the European DW study, three surveys were rolled out, two contained 15 PC questions each and differed on whether the conditions were chronic or temporary, the last version had five PC questions[9,18].

Step 3: Population health equivalence questions and analysis

PHE questions were used to anchor disability weights from 0 (perfect health) to 1 (death). In them, respondents were asked to choose, in hindsight, between two programs that had produced the greatest overall population health benefit. The first program prevented 1,000 people from getting an illness that caused rapid death, while the second program prevented a higher number of people, also referred to as bids, (between 1,100 to 10,000 depending on the randomization procedure) to get a non-fatal illness that causes lifelong health problems of a health state. Each time a PHE question was asked, the health state that was prevented would be randomly chosen from a pool of 30 health states chosen to represent a range from mild to severe health states from the GBD 2010 DW study. Although PHE data was collected in the European DW study, the authors excluded these for their estimation of European DW[18] due to a lack of consistency and variation in the responses. Hence, in the estimation GBD 2013 DW, Salomon et al. relied only on PHE data from the GBD 2010 DW study[10].

PHE data were analyzed through an interval regression in logit transformed space specified as:

$$left,\ right = \beta_1 seqdum1 + \cdots + \beta_{30} seqdum30 + \varepsilon$$

Where:

- left is the logit transformed DW estimate when respondent chose the program that prevented illness. If a respondent indicated that a program that averted 3,000 people getting a stroke is better than a program that averted 1,000 people dying, then this indicates that the respondent attached a DW of 0.33 or higher. In other words, the left limit is 0.33.
- Right is the logit transformed DW estimate when respondent chose the program that prevented rapid death: If a respondent indicated that a program that averted 3,000 people

getting a stroke is not as good as a program that averted 1,000 people dying, then this indicates that the respondent attached a DW of 0.33 or lower. In other words, the right limit is 0.33.

- Seqdum `i` either 0 or 1. i denotes a different health state that was asked for the program that prevents disease.

In both the GBD 2010 DW study and the European DW study, only one of the survey versions contained three PHE questions. Importantly, the versions that contained the PHE questions in the 2010 study were only rolled out in the Web sample[9,18].

Step 4: Anchoring results from paired comparison to 0 to 1 disability weights

An OLS regression was run with the following specification:

$$\text{pheparam} = \beta_0 + \beta_1 probitparam + \varepsilon$$

Where pheparam are the coefficients from the interval regression, and probit param are the coefficients from the probit regression. After running this OLS regression to anchor the PC responses between 0 and 1, a prediction of *pheparam* was stored based on the model, by taking β_0 and β_1 random samples from its multinomial distribution and using its variance-covariance matrix:

$$\hat{\theta}_{global} = E[pheparam|probitparam] = \beta_0 + \beta_1 probitparam$$

Step 5: Survey specific adjustments to disability weights estimates

To get survey-specific standard deviation adjustments, the standard deviation of the residual was estimated, stored and used in 10,000 simulations to incorporate variability to the DW estimates.

$$\hat{\theta}_{global} = \beta_0 + \beta_1 probitparam_{survey} + \varepsilon$$

$$residual = prediction - \hat{\theta}_{global}$$

The mean of the inverse logit of the 10,000 simulated values of $\hat{\theta}_{global}$ were then saved as the DW for each health state.

Step 6: Estimating uncertainty

Finally, to estimate the uncertainty interval per each DW, 1,000 samples of the data were retrieved and 1,000 DW were estimated per health state. Each DW were inverse-logit transformed, the mean

of the samples became the DW per health state, while the uncertainty was derived from the 2.5^{th} and 97.5^{th} bootstrapped sample percentiles[9,10,19,20].

Health valuation and respondent characteristics

The previous paragraphs summarize the evolution and methodology of disability weights in the GBD study leaving one fundamental topic out. In Salomon et al. 2010, the starting point in the estimation of disability weights was defining whether health or wellbeing loss should be measured. Before 2010, the distinction between both concepts was not clearly established. In fact, the GBD DW study of 2010 explicitly stated that health and not wellbeing loss would be measured. One of the reasons why health loss was preferred over wellbeing loss was that evaluating health by considering failures in the body and mental functioning should, in theory, be universal and less dependent of context. On the contrary, questions about loss of wellbeing could be highly dependent on the common values of different populations[9].

The GBD 2010 study and the European DW study rolled out across different regions of the world proved that there was indeed very little difference in how people from countries as different as Tanzania and the Netherlands valuated health[9,18]. Salomon et al. 2015 found high correlation between the health state preferences of paired comparison question across countries in all data included in the GBD 2013 DW study. However, two different DW surveys from 2021 onwards in Japan and China have shown that there are more differences than expected in how people value health across countries. Using data from Japan, Nomura et al. 2021 estimated DW two to three times larger than GBD 2013 disability weights for 20 of the health states valuated, and two to three times smaller than GBD 2013 for 23 of the health states valuated. When studying the reasons for the differences, Nomura et al. looked closer at the lay descriptions of health loss and concluded that Japanese people gave lower DWs to mental symptoms and substance use, and higher DWs to health states that included sensory and pain symptoms[19]. In China, Liu et al. 2022 also found considerably lower disability weights for mental disorders, alcohol use and dementia than those in GBD 2013[20].

The empirical differences between new disability weights following the GBD methodology suggest that measuring health loss instead of wellbeing may not necessarily translate into universal health loss valuation. Moreover, it raises the question whether differences across individual characteristics and disability weights have been thoroughly examined across individual personal characteristics such as sex or age.

Previous literature on individual characteristics and their association to disability weights following the GBD methodology have focused on the differences of paired comparison questions. In the European DW study Maertens de Noordhout et al. 2018 studied whether sex, age, education, disease experience, and income affects health loss valuation using the paired comparison data. They found correlation estimates of health preferences to be above 0.97 between different groups and concluded that probit coefficients did not vary widely across characteristics[22]. Similarly, in the Chinese DW study of 2022, Liu et al. found high correlation between paired comparison probit coefficients across age, education, income, sex, disease status, medical background and professions[20]. In the latest Dutch disability weights study from 2024, Haagsma et al. found respondent characteristics had no influence on health state valuation with paired comparison data[23].

Although the evidence suggests that respondent characteristics do not affect health loss valuation in the estimation of disability weights, previous literature has been limited to the analysis of paired comparison questions. There is little to no discussion on differences of sex disaggregated disability weights differences due to population health equivalence data.

Purpose of this study

Analyze differences by sex of the respondent in the disability weights used for the Global Burden of Disease study.

Methods

Data

Our study uses data from the GBD 2013 disability weights study, which pooled together the GBD DW 2010 study and the European DW study.

Statistical analysis

First, we compared respondents' characteristics to national level estimates by using GBD study estimates on the distribution of age, sex and education in the years and countries the disability weights surveys were conducted. Then, to evaluate whether health loss valuation in GBD varies by sex, we stratified both PC and PHE data by sex and estimated three sets of sex disaggregated disability weights. The first set of disability weights used sex disaggregated PC data and all PHE data,

the second set used all PC data and sex disaggregated PHE data, the third set used sex disaggregated PC and sex disaggregated PHE data. In other words, our statistical analysis followed disability weights steps 5 to 6 described above, but with sex stratified PC and PHE data in the estimation of PC probit regression and PHE interval regression.

We analyzed two plots to visually test the consistency of PC and PHE data stratified by sex. The first one is a heat map of the health state pairs elicited in the PC questions where the probability of choosing health state A over health state B is depicted by color. We checked whether health states with higher probabilities of being chosen as healthier slowly decrease their probability of being chosen as the health states they are compared to decrease in severity. To visually vet the quality of PHE data we ordered health states from least to most severe and from smallest to largest bids and plot the probability of choosing the program that averts illness over premature death. We visually checked for the probability of choosing the second program to increase monotonically as the bids increase per health state, and as the health states increase in severity. Both approaches followed the methodology established in the GBD 2010 DW study and the European study[9,10,18].

We then tested whether the differences in DW by sex are statistically significant, by estimating relative and absolute DW differences for each pair of DW in 1000 bootstrapped draws. We determined that the differences found were statistically significant if the 95% uncertainty interval derived for the relative difference did not cross 1, or 0 for the absolute difference. Finally, we estimated the impact of sex disaggregated differences in YLDs. In the GBD study, YLDs are estimated for health outcomes that follow four levels of hierarchical structure depending on the level of detail, with level 1 being the most aggregated hierarchy[21]. Our analysis used the level 3 health causes and selected only the 20 causes with the highest YLD estimates for the GBD 2023 study. We then selected a subsample of health states that used sequelae for which disability weights were directly derived from the GBD 2013 DW study. These sequelae were lastly multiplied by the female to male disability weights ratio and aggregated at the cause level to derive YLDs with female only disability weights. The equivalent conversion was done for the estimation of YLDs with male only disability weights.

Finally, we vetted our PC-PHE sex disaggregated disability weight results by comparing disability weight estimates across health state severity. We reviewed, for example, whether mild, moderate and severe anemia disability weights increased for both sets of male and female disability weights estimated.

Results

As is evident in Table 2, we observed that the GBD 2013 DW study survey had similar percentages of respondents in different age categories when compared to national estimates of individuals 18 years old and older, except for the United States which was slightly skewed towards respondents older than 50 years old. The sex distribution across survey respondents was also similar to national estimates, except for the United States which had 63 percent female respondents when the national estimate was 51 percent. Respondents in all countries had higher education levels than the country level estimates, except for Italy, the Netherlands and Sweden where the proportion of survey respondents with less than 12 years of education or at 12 or 13 years of education was higher than the national estimates. Respondents from 167 countries answered PHE questions in the GBD web survey. The majority were female (70 percent) and had more than 13 years of schooling (94 percent). Figure 1 shows the distribution of respondents per country of the 2010 internet survey. Most respondents were from the United States, Australia and Western Europe; there was very little representation from Africa, Caribbean and Central America, and Asia.

Table 2. Respondent's characteristics of GBD 2013 disability weights study

	Web (PC)	Web (PHE)	BGD (survey)	BGD	IDN (survey)	IDN	PER (survey)	PER	TZA (survey)	TZA	USA (survey)	USA	ITA (survey)	ITA	HUN (survey)	HUN	NLD (survey)	NLD	SWE (survey)	SWE
Respondents	15,923	11,129	2,610		2,430		2,925		2,613		3,054		8,054		6,053		8,005		8,548	
PC questions	179,522		39,150		36,450		42,834		38,786		42,751		96,590		72,705		96,215		102,690	
PHE questions		33,296																		
Sex																				
Males	32.2	29.5	57.7	49.8	51.7	49.8	52.7	49.6	57.7	47.6	36.6	48.6	48.3	49.5	47.7	49.3	48.0	50.2	48.0	50.8
Females	67.6	70.3	42.3	50.2	48.3	50.2	47.3	50.4	42.3	52.4	63.4	51.5	51.8	50.5	52.3	50.8	52.0	49.8	52.0	49.2
Missing	0.2	0.2																		
Age																				
18-29	31.7	30.8	31.3	31.5	27.9	27.8	32.2	28.9	37.4	37.2	5.3	19.2	19.4	17.7	23.5	20.0	16.4	20.6	23.8	23.4
30-39	24.0	23.4	23.8	25.3	28.8	26.0	24.5	24.0	22.6	26.3	10.9	17.9	25.1	22.7	21.9	25.1	19.7	20.2	19.6	21.7
40-49	16.7	17.1	20.2	18.5	22.8	20.6	21.8	18.5	17.5	16.4	15.7	19.3	23.8	26.9	21.8	22.0	26.8	25.4	23.5	23.7
50+	27.6	28.7	24.8	24.7	20.5	25.7	21.5	28.6	22.5	20.2	67.9	43.6	31.7	32.8	32.7	33.0	37.2	33.7	33.1	31.3
Missing											0.3									
Education																				
None	0.3	0.0	43.1	47.5	6.6	15.0	1.1	12.4	35.1	36.1	0.2	1.0	0.1	1.0	0.0	0.6	0.8	0.7	0.4	0.1
< than 12y	0.4	0.5	31.5	46.5	40.0	67.6	10.3	62.5	26.4	59.8	7.1	14.4	39.4	39.5	38.7	47.0	33.9	18.2	13.8	13.2
12 or 13y	5.9	5.5	21.6	3.1	18.2	11.3	47.1	4.5	37.4	1.8	29.0	35.9	41.7	14.4	39.8	31.2	35.0	14.6	50.6	18.3
> than 13y	93.1	93.7	3.8	3.0	35.3	6.1	41.5	20.6	1.1	2.3	63.4	48.8	18.5	45.1	21.5	21.3	30.0	66.4	34.1	68.4
Missing	0.3	0.2									0.4		0.3		0.0		0.3		1.2	

National estimates for face-to-face and telephone GBD 2010 surveys, as well as European web surveys are depicted next to survey characteristics.

Figure 1. Web survey respondents in 2010

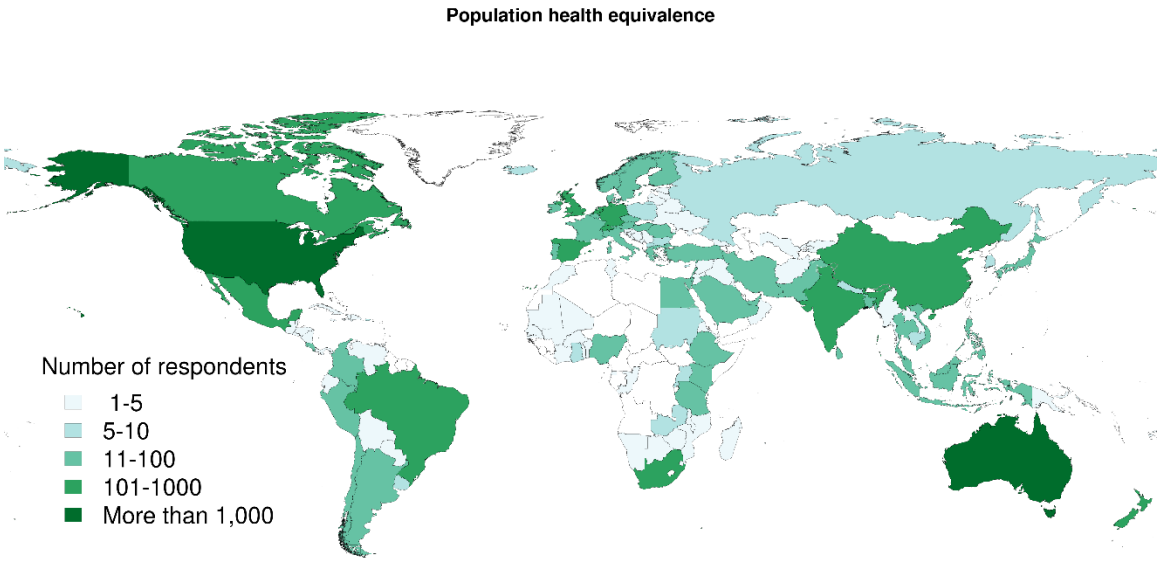


Figure 2 shows the consistency of the sex specific PC samples. The probability of choosing the first health state over the second health state is shown in different colors, the x axis represents the first option, and the y axis represents the second health state option. This is a matrix of 367 by 367 health states, representing all the combinations of health states asked in the paired comparison questions. The smooth transition from blue to red indicates overall consistency in which health states were preferred in relation to other health states for both female and males. When we disaggregated PHE data by sex, we observed greater discrimination between health states and bids for female responses in comparison to males. In Figure 3, health states are ordered from least to most severe and each dot per health state represents a bid or the number of people who would benefit from the program that averts disease. Females generally were more likely to choose 'cure of a health state' over preventing death with each successive increase in the 'bid' among a health state, and severity of the health state.

Figure 2. Response probabilities in paired comparison questions in GBD 2013 DW study

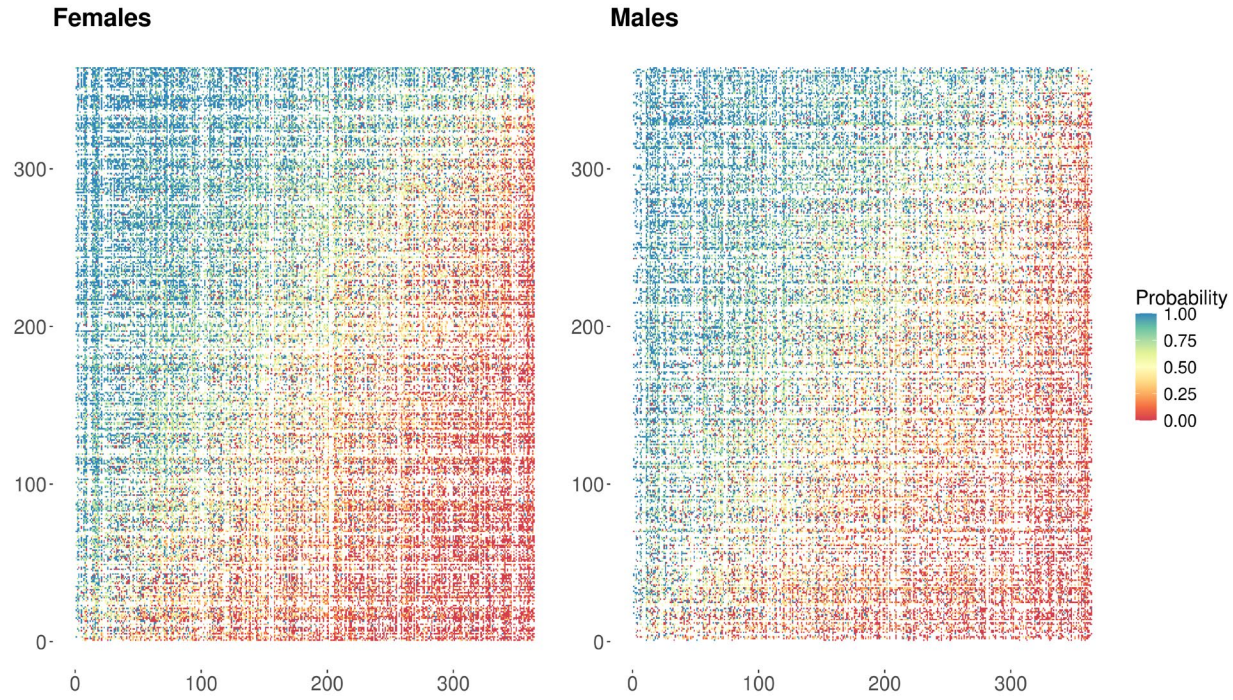
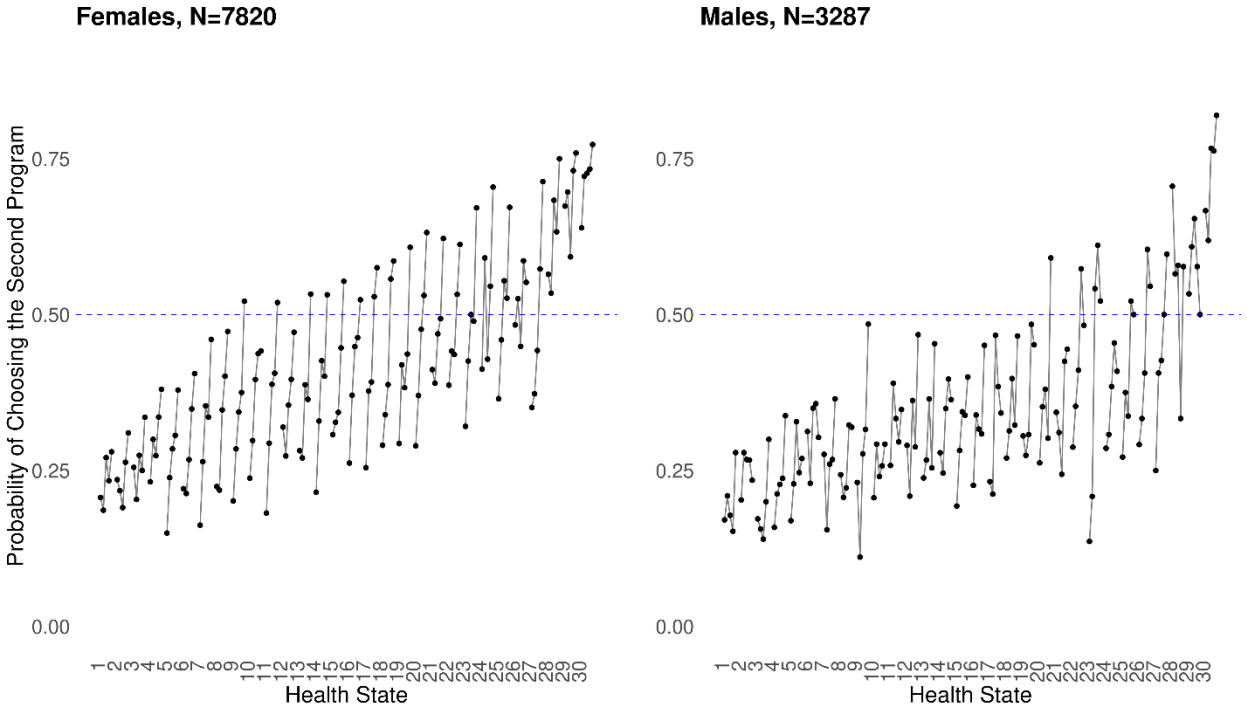


Figure 3. Response probabilities in population health equivalence questions in GBD 2010 DW study



Moving on, we compared male and female disability weights in multiple scenarios where we varied whether we stratified the PC data, PHE data or both. Coefficients of the probit regression on the PC data by sex were highly correlated, with a Pearson's R of 0.985 (95% CI 0.982 – 0.988). This high correlation is noted in Figure 4 upper left panel, which shows the estimated disability weights when only the paired comparison data is stratified by sex. Significant differences between male and female disability weights are denoted with a yellow star symbol. Only one health state was significantly different between the sexes. Females' estimated disability weight for anorexia nervosa was 1.98 (95% UI 1.114 – 3.185) times the disability weight estimated for males. Table 3 shows a list of disability weights with the 10 highest female to male ratio, and the 10 highest male to female ratios.

The upper right panel in Figure 4 shows the comparison between males and females disability weights when only the population health equivalence data is stratified by sex. This plot nicely illustrates the application of PHE data in the disability weights methodology. PHE interval regression coefficients are used to anchor the probit regression preferences from 0 to 1, therefore higher female PHE interval coefficients shift the disability weights upwards in comparison to male disability weights. Appendix 1 shows this is a parallel shift in logit space between PHE interval regression and PC probit regression coefficients. In other words, the order of the preferences of health state does not change, instead all the estimated female disability weights shift upward while keeping the same severity rank as males.

The left lower panel in Figure 4 presents estimated disability weights with PC and PHE sex disaggregated data. Meaningful and significant differences between female only and male only disability weights can be observed, with every disability weight estimated with female only data being higher than males. The largest absolute differences are for Bipolar disorder (maniac episode), Terminal phase, without medication (for cancers), and AIDS without ARV treatment, where females disability weight are 2.92 (95% UI 1.36-6.45), 2.39 (95% UI 1.18-5.17), and 2.33 (95% UI 1.18-5.17) times male disability weights respectively. Appendix 2 presents the top 10 estimated DW for females and males ordered from greatest to least absolute differences. These disability weights differences translate into higher YLDs when using female only disability weights versus the current and male only disability weights. Figure 5 shows what YLDs would be if we used sex specific DW as well as the current sequalae exposure in the GBD 2023 study. YLDs with female disability weights are always higher than the current ones and YLDs with male disability weights are always lower than the current ones. For example, for asthma, YLDs using female only DWs are 205.99 per 100,000 higher than

those using male only DWs. For alcohol use disorders and bipolar disorders, the differences are 153.74 per 100,000, and 75.71 per 100,000, respectively.

Figure 4. Sex disaggregated disability weights with GBD data

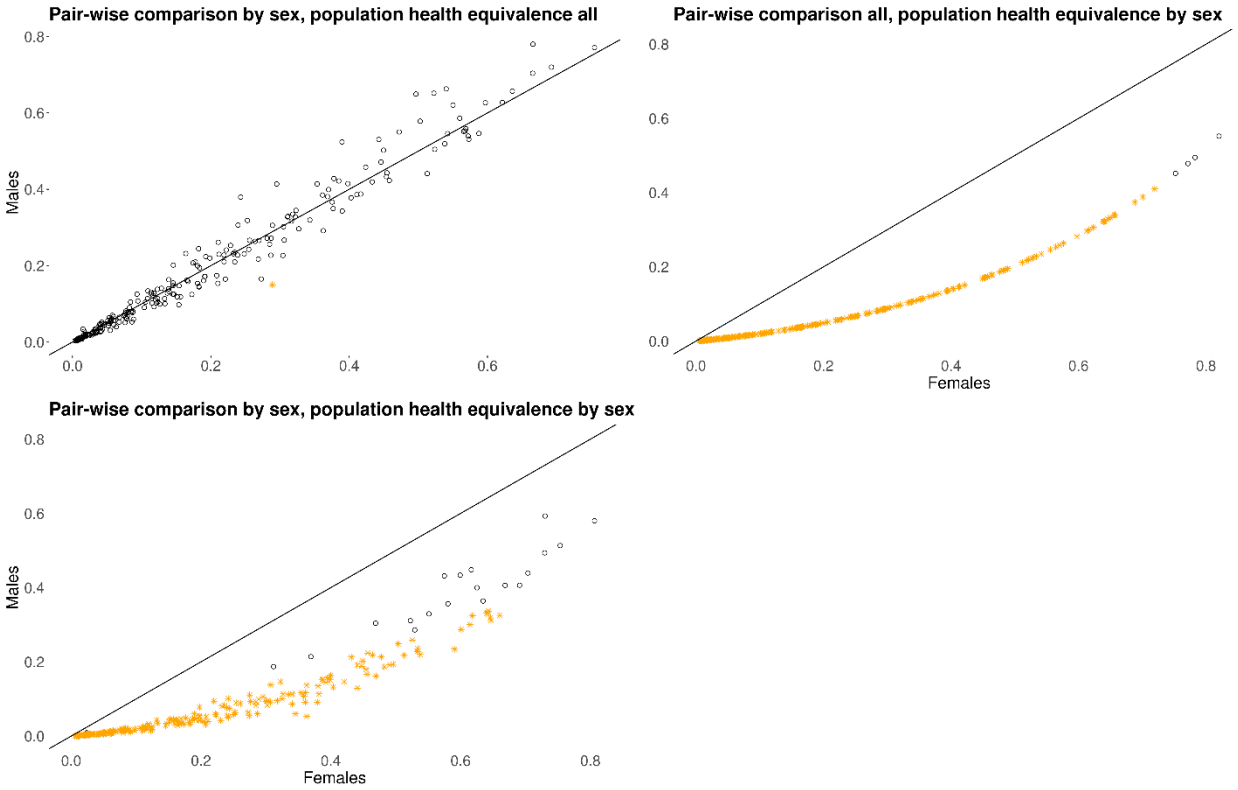
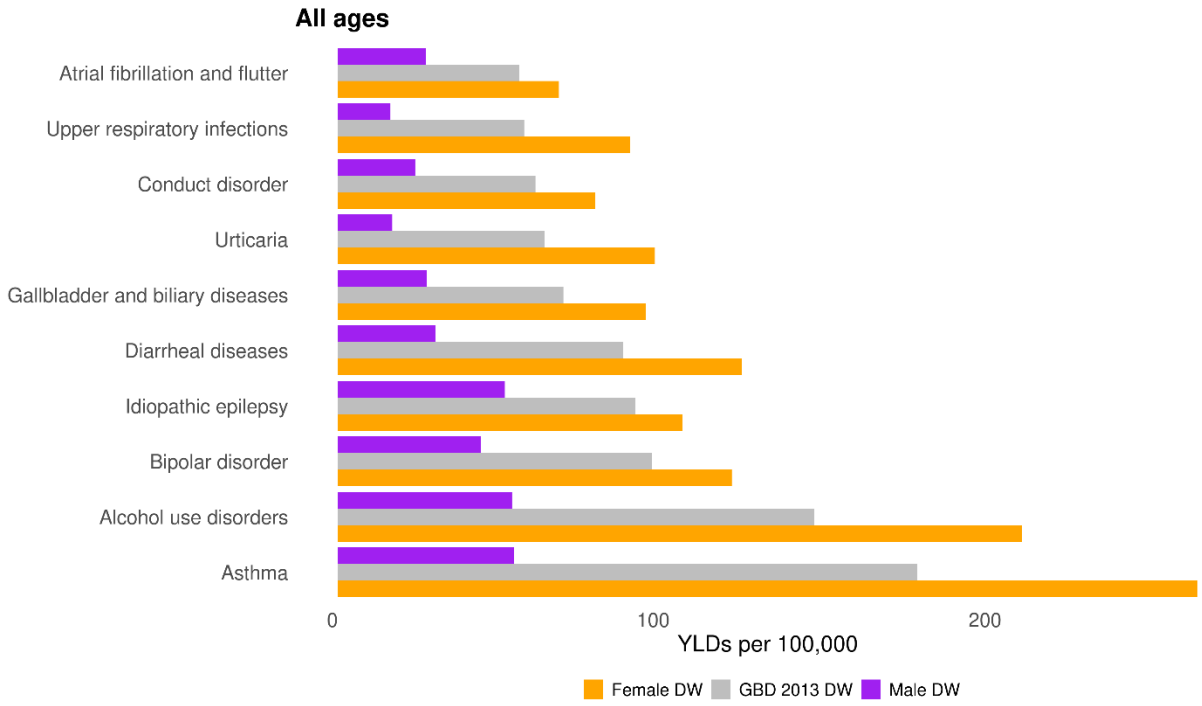


Table 3. Sex disaggregated PC and both sexes PHE disability weights

Health state	Female DW	Male DW	GBD DW	Female:male ratio
Female > Male				
Amputation of finger(s), excluding thumb	0.007 (0.003-0.014)	0.004 (0.001-0.009)	0.005 (0.002-0.011)	2.065 (0.519-5.671)
Amputation of toe	0.008 (0.003-0.017)	0.005 (0.002-0.01)	0.006 (0.003-0.013)	2.035 (0.544-5.359)
Anorexia nervosa	0.289 (0.19-0.399)	0.15 (0.103-0.209)	0.227 (0.154-0.314)	1.997 (1.16-3.252)
Parkinson's disease, mild	0.014 (0.007-0.025)	0.008 (0.004-0.015)	0.011 (0.005-0.02)	1.953 (0.664-4.38)
Cannabis dependence, mild	0.055 (0.033-0.084)	0.031 (0.018-0.048)	0.042 (0.026-0.064)	1.919 (0.885-3.61)
Anemia, mild	0.005 (0.002-0.011)	0.003 (0.001-0.007)	0.004 (0.001-0.009)	1.826 (0.411-5.49)
Other injuries of muscle and tendon (includes sprains, strains and dislocations other than shoulder, knee, hip)	0.01 (0.004-0.019)	0.006 (0.003-0.013)	0.008 (0.003-0.017)	1.815 (0.555-4.644)
Kwashiorkor	0.063 (0.038-0.095)	0.04 (0.02-0.069)	0.053 (0.032-0.082)	1.768 (0.741-3.438)
Infectious disease, acute episode, mild	0.008 (0.003-0.016)	0.005 (0.002-0.011)	0.006 (0.003-0.013)	1.765 (0.493-4.741)
Bulimia nervosa	0.273 (0.172-0.379)	0.165 (0.113-0.226)	0.225 (0.151-0.315)	1.712 (0.973-2.728)
Male > Female				
Stress incontinence	0.015 (0.008-0.028)	0.034 (0.02-0.053)	0.021 (0.012-0.037)	0.48 (0.193-1.028)
Spinal cord lesion below neck level (treated)	0.243 (0.153-0.344)	0.379 (0.267-0.517)	0.299 (0.203-0.41)	0.661 (0.384-1.056)
Impotence	0.017 (0.008-0.03)	0.028 (0.015-0.046)	0.019 (0.01-0.033)	0.664 (0.263-1.4)
Distance vision blindness	0.164 (0.106-0.23)	0.232 (0.163-0.313)	0.19 (0.133-0.26)	0.727 (0.42-1.145)
Intellectual disability, severe	0.295 (0.188-0.415)	0.414 (0.259-0.591)	0.325 (0.216-0.452)	0.748 (0.399-1.302)
Gastric bleeding	0.146 (0.092-0.208)	0.202 (0.134-0.279)	0.165 (0.112-0.226)	0.752 (0.406-1.265)
Dementia, severe	0.39 (0.253-0.525)	0.524 (0.373-0.677)	0.445 (0.302-0.587)	0.761 (0.47-1.167)
Motor plus cognitive impairments, moderate	0.182 (0.116-0.254)	0.244 (0.162-0.338)	0.207 (0.138-0.289)	0.772 (0.439-1.25)
Intellectual disability, moderate	0.093 (0.058-0.136)	0.125 (0.082-0.184)	0.104 (0.069-0.149)	0.778 (0.404-1.378)
Stroke, long-term consequences, severe	0.497 (0.323-0.653)	0.649 (0.468-0.81)	0.545 (0.381-0.698)	0.78 (0.481-1.213)

Figure 5. GBD 2023 global YLDs estimated with PC and PHE sex disaggregated disability weights



Finally, we observed that our PC-PHE sex disaggregated disability weights pass the face validity test. Appendix 3 shows that for almost every health state, disability weights increase as the severity of conditions increases. Only in three health conditions does the disability weights not increase monotonically with severity. For female disability weights, severe distance vision disability weight is higher than blindness, and severe hearing loss with ringing has a higher disability weight than complete hearing loss with ringing. For male disability weights, severe stroke long-term consequences has a higher disability weight than severe plus cognition stroke long term-consequences.

Discussion

We found that when we only disaggregate paired comparison data by sex of the respondent, there is one significantly different disability weight between females and males from 235 health states used in the GBD 2013 study. Anorexia nervosa is viewed as more severe by females than males. When only disaggregating population health equivalence questions, we found that female disability weights would all be higher than those for males. These differences are reflected in the estimation of disability weights by sex when both PC and PHE data are disaggregated. If disability weights respondents were only female then all disability weights would be higher than the ones currently in

use, conversely if respondents were only males, then all disability weights would be lower than the current set.

Comparing disability weights estimated with sex disaggregated data is a small part in understanding whether females and males value health loss differently. Greater insight is gained when contrasting our findings to the disability weights methodology. Paired comparison questions are the primary input to elicit health state preferences, and PHE questions are used to transform these preferences to an appropriate scale for the estimation of YLDs. The high correlation between sex disaggregated PC probit coefficients and the large differences in DW estimates when disaggregating PHE data suggest that preferences for health states do not differ by sex of the respondent, but that females and males do differ on their willingness to accept loss of life or disability as a health program evaluator.

Previous literature found similarly high levels of correlation between female and males paired comparison probit coefficients. The 0.985 Pearson correlation estimate in our analysis is similar to 0.997 found in the Chinese disability weights study of 2020, and to 0.978 in the Dutch disability weights study of 2025[20,23]. However, we did not find literature quantifying how health valuation differed by sex of the respondents when using PHE questions or the impact of these differences in the estimation of sex disaggregated disability weights. To our knowledge, this is the first study that quantifies the difference in disability weights when PHE data is disaggregated by sex.

Our conclusions are limited to the scope of our input data. As presented in Table 1 and Figure 1, most of the data in the estimation of DW were collected in Western Europe, the United States, Australia and a few other high-income countries. The only explicit effort to include low- and middle-income countries was in the inclusion of household surveys in the 2010 study. However, the data provided from these surveys were only on paired comparison questions, meaning the findings from the PHE data are constrained to the internet survey. Although we were able to analyze the sex differences in disability weights used by the GBD, we caution against making greater generalizations due to lack of representation in Africa, Eastern Europe and Latin American. We also acknowledge that this study only describes the differences by sex without consideration of other demographic variables that could explain these differences. The next chapter addresses this gap and studies whether the sex differences in the PHE data hold when accounting for age and education.

Bibliography

1 Mathers CD, Murray CJ, Ezzati M, Gakidou E, Salomon JA, Stein C. Population health metrics: crucial inputs to the development of evidence for health policy. *Population Health Metrics* 2003; **1**: 6.

2 Murray CJ, Salomon JA, Mathers C. A critical examination of summary measures of population health. *Bull World Health Organ* 2000; **78**: 981–94.

3 Rosenberg MA, Fryback DG, Lawrence WF. Computing Population-based Estimates of Health-adjusted Life Expectancy. *Med Decis Making* 1999; **19**: 90–7.

4 Kaplan RM, Erickson P. Gender differences in quality-adjusted survival using a Health-Utilities Index. *American Journal of Preventive Medicine* 2000; **18**: 77–82.

5 Murray CJ, Lopez AD. Quantifying disability: data, methods and results. *Bull World Health Organ* 1994; **72**: 481–94.

6 World Health Organization, Baltussen, Rob M. P. M, Adam, Taghreed, Tan-Torres Edejer, Tessa, Hutubessy, Raymond C. W. et al. Making choices in health : WHO guide to cost-effectiveness analysis. World Health Organization. 2003. https://iris.who.int/handle/10665/42699

7 Bill & Melinda Gates Foundation, NICE International, Health Intervention and Technology Assessment Program. Methods for Economic Evaluation Project (MEEP). 2014. https://www.idsihealth.org/wp-content/uploads/2015/01/MEEP-report.pdf

8 Vos T, Lim SS, Abbafati C, *et al*. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* 2020; **396**: 1204–22.

9 Salomon JA, Vos T, Hogan DR, *et al*. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *The Lancet* 2012; **380**: 2129–43.

10 Salomon JA, Haagsma JA, Davis A, *et al*. Disability weights for the Global Burden of Disease 2013 study. *Lancet Glob Health* 2015; **3**: e712-723.

11 Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health: an international perspective. In: Brazier J, Ratcliffe J, Saloman J, Tsuchiya A, eds. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press, 2016: 0.

12 Haagsma JA, Polinder S, Cassini A, Colzani E, Havelaar AH. Review of disability weight studies: comparison of methodological choices and values. *Population Health Metrics* 2014; **12**. DOI:10.1186/s12963-014-0020-2.

13 Essink-Bot ML, Bonsel GJ. How to derive disability weights. In: Summary measures of population health : concepts, ethics, measurement and applications. World Health Organization, 2002 https://apps.who.int/iris/handle/10665/42439 (accessed April 25, 2023).

14 Salomon JA, Murray CJL, Ustun B, Chatterji S, Health state valuation in summary measures of population health. In: Health systems performance assessment : debates, methods and empiricism. World Health Organization. 2003. https://www.who.int/publications/i/item/9241562455 (accessed June 8, 2025).

15 Murray CJL, Salomon JA, Mathers CD, Lopez AD. Summary measures of population health: Conclusions and recommendations. In: Summary measures of population health : concepts, ethics, measurement and applications. World Health Organization, 2002 https://iris.who.int/handle/10665/42439 (accessed June 8, 2023).

16 Murray CJL, Lopez AD, Organization WH, Bank W, Health HS of P. The Global burden of disease : a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020 : summary. World Health Organization, 1996 https://apps.who.int/iris/handle/10665/41864 (accessed April 26, 2023).

17 Cj M, Ad L. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: global Burden of Disease Study. *Lancet (London, England)* 1997; **349**. DOI:10.1016/S0140-6736(96)07494-6.

18 Haagsma JA, Noordhout CM de, Polinder S, *et al*. Assessing disability weights based on the responses of 30,660 people from four European countries. *Population Health Metrics* 2015; **13**. DOI:10.1186/s12963-015-0042-4.

19 Nomura S, Yamamoto Y, Yoneoka D, *et al*. How do Japanese rate the severity of different diseases and injuries?—an assessment of disability weights for 231 health states by 37,318 Japanese respondents. *Popul Health Metrics* 2021; **19**: 21.

20 Liu X, Wang F, Zhou M, *et al*. Eliciting national and subnational sets of disability weights in mainland China: Findings from the Chinese disability weight measurement study. *The Lancet Regional Health – Western Pacific* 2022; **26**. DOI:10.1016/j.lanwpc.2022.100520.

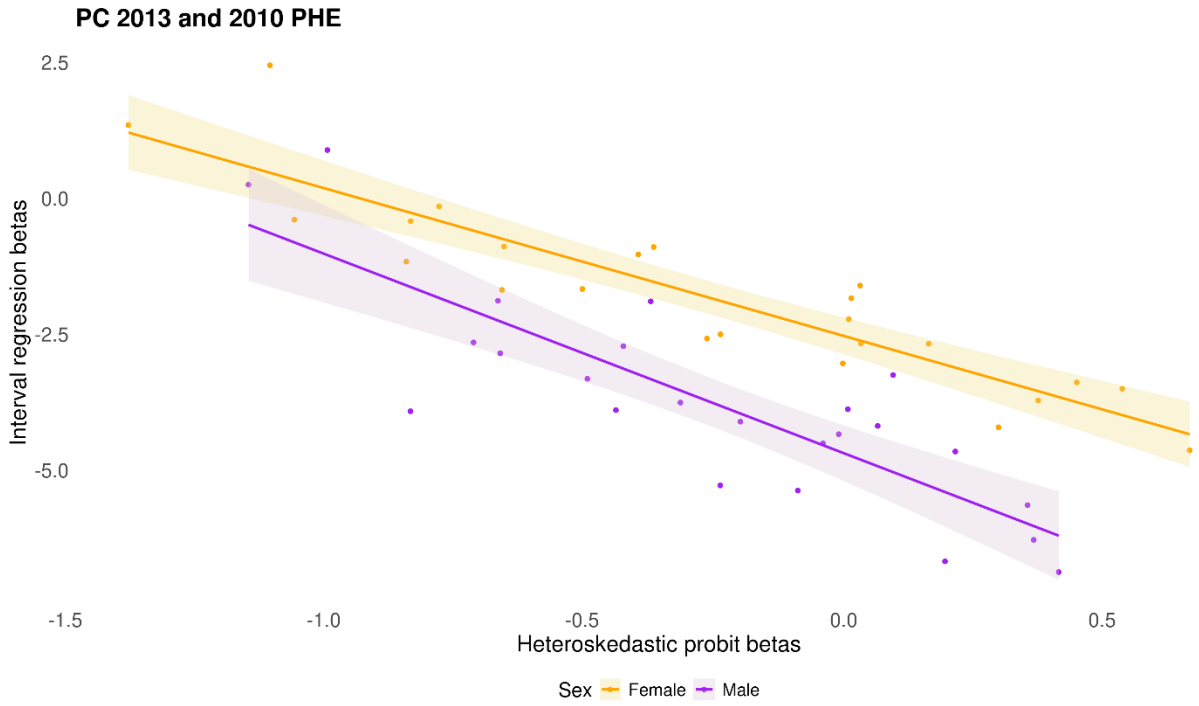
21 Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021 - The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00933-4/fulltext (accessed June 8, 2025).

22 Maertens de Noordhout C, Devleesschauwer B, Salomon JA, *et al*. Disability weights for infectious diseases in four European countries: comparison between countries and across respondent characteristics. *Eur J Public Health* 2018; **28**: 124–33.

23 Haagsma JA, Charalampous P. Deriving disability weights for the Netherlands: findings from the Dutch disability weights measurement study. *Popul Health Metrics* 2024; **22**: 26.

Appendix

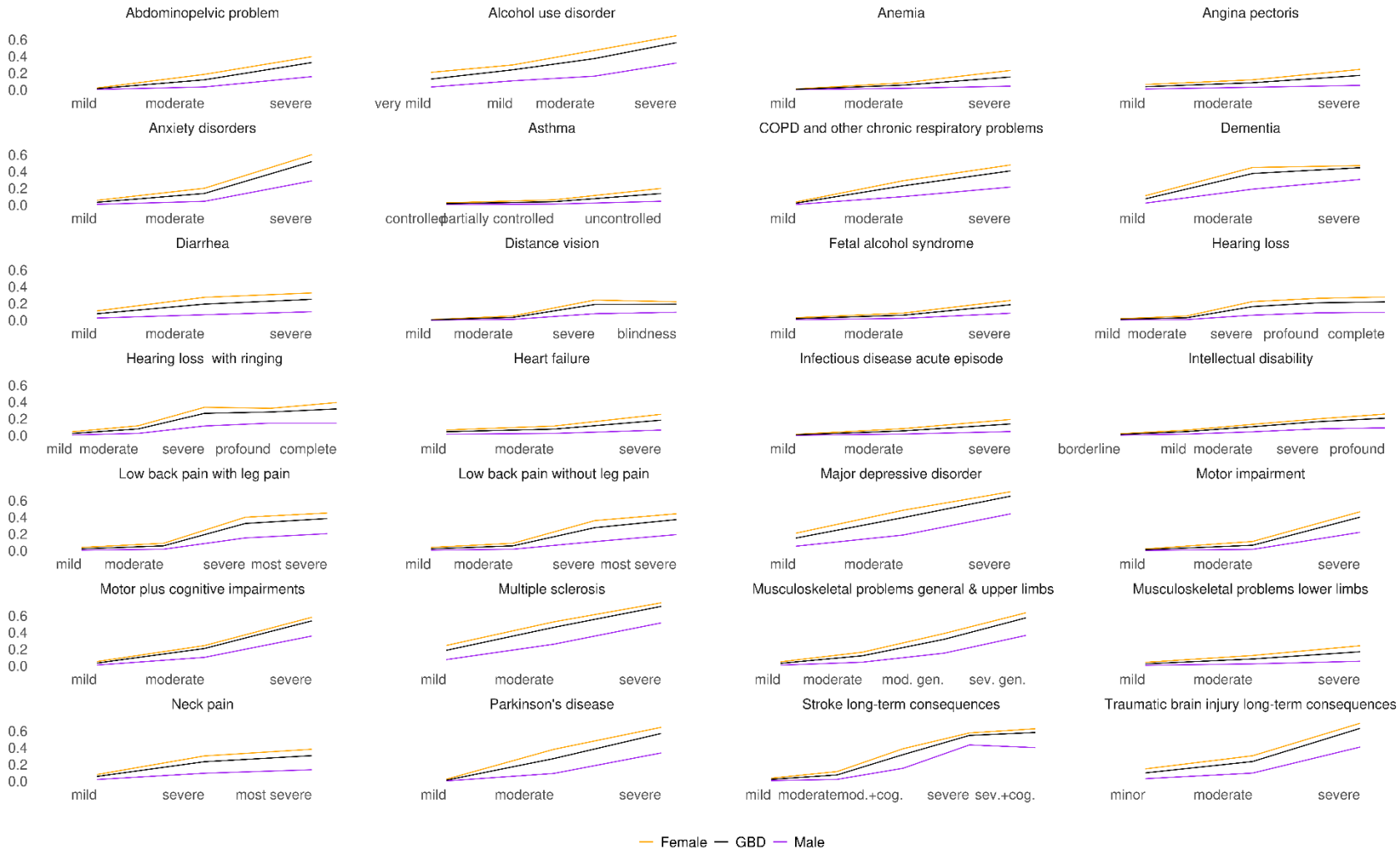
Appendix 1. Probit paired comparison coefficients against interval regression population health equivalence coefficients



Appendix 2. Top 10 absolute differences between female and male PC and PHE sex disaggregated disability weights

Health state	Female DW	Male DW	GBD DW	Female:male ratio
Female > Male				
Bipolar disorder, manic episode	0.591 (0.419-0.725)	0.234 (0.095-0.419)	0.489 (0.339-0.632)	2.924 (1.356-6.446)
Terminal phase, without medication (for cancers, end-stage kidney/liver disease)	0.647 (0.469-0.789)	0.311 (0.122-0.525)	0.562 (0.392-0.713)	2.386 (1.181-5.166)
AIDS cases, not receiving ARV treatment	0.66 (0.48-0.795)	0.325 (0.134-0.549)	0.574 (0.401-0.724)	2.328 (1.149-5.241)
Alcohol use disorder, severe	0.646 (0.469-0.779)	0.319 (0.134-0.551)	0.563 (0.396-0.717)	2.322 (1.131-5.049)
Amputation of both lower limbs (long term, without treatment)	0.538 (0.381-0.686)	0.22 (0.093-0.401)	0.44 (0.294-0.584)	2.838 (1.28-6.257)
Terminal phase, with medication (for cancers)	0.614 (0.444-0.749)	0.301 (0.124-0.514)	0.534 (0.368-0.683)	2.353 (1.123-5.167)
Anxiety disorders, severe	0.601 (0.434-0.735)	0.287 (0.115-0.499)	0.518 (0.356-0.665)	2.406 (1.156-5.311)
Vesicovaginal fistula	0.441 (0.302-0.568)	0.129 (0.049-0.251)	0.341 (0.232-0.469)	4.054 (1.638-9.407)
Anorexia nervosa	0.363 (0.254-0.476)	0.053 (0.022-0.102)	0.227 (0.154-0.314)	7.929 (3.281-17.27)
Disfigurement, level 3, with itch/pain	0.642 (0.465-0.777)	0.332 (0.13-0.566)	0.569 (0.392-0.722)	2.239 (1.047-5.054)

Appendix 3. PC and PHE sex disaggregated disability weights across health state severity



Chapter 3: Differential health loss valuation by sex on population health equivalence questions

Introduction

Overview

The last chapter clearly showed that differences by sex of the respondent in disability weights used in the GBD study are due to differences in health valuation of population health equivalence (PHE). In this chapter we examine whether the PHE differences by sex hold when we account for other individual characteristics such as age and education. We start by discussing the PHE method, its importance in health valuation for global health, and the literature on whether sex of the respondents affects PHE health valuation. We then define the statistical method we used to accomplish our research purpose, present our results, and finally reflect on the implications of our findings in the GBD study.

Trade off methods

Health valuation and its measurement follows a long and rich literature in health economics, psychology, psychometrics, sociology, statistics and philosophy[1]. Population health equivalence questions (PHE) developed by Salomon et al. 2010 stem from person trade-off methods (PTO) described and developed by Nord[2] and used in the 1996 GBD disability weights study[3]. Together with time trade-off (TTO) and standard gamble (SG), PTO and PHE fall under the family of trade-off health valuation methods in which respondents are asked to give up something valuable in order to gain health[4]. In standard gamble, respondents need to choose between living in a chronic health state for a number of years or try a new option that could result either in perfect health or death[5]. In time trade-off, respondents are asked how much time they would be willing to give up in order to avoid living with a health condition[6]. Person trade-off questions ask respondents to choose between investing in population level interventions that either extend the life of people living in ideal health, versus extending the life of people living in less than ideal health[3,7]. Similarly to PTO, PHE questions ask respondents to choose between two population health programs but avoid resource allocation

decisions to help respondents focus on health constructs instead of wellbeing, and phrase questions retrospectively so respondents would not feel responsible for life-and-death consequences[3].

Of these trade-off methods, SG and TTO were regarded as the most common methods used to valuate health. SG was often used because of theoretical underpinnings, and TTO because of its application in summary measures of population health[6,8,9]. The standard gamble is rooted in expected utility theory, and as such considered by few as the closest to the gold standard in health valuation[10]. The time trade-off method is used in the estimation of QALYs, making it the focus of European studies[6,9,11–13]. Since the first Global Burden of Disease in 1994, DALYs have taken a greater international role as both a measure of disease burden and a metric for cost-effectiveness analysis, making the current GBD disability weights methods, along with the PHE method, one of the most common methods to valuate health since 2010[14,15].

Trade off methods and respondent characteristics

Despite their wide use, there is no published research on the differences between females and males' valuation of health using PHE questions. This may be due to its use as an intermediate anchoring step of health state preferences elicited by paired comparison questions when calculating disability weights rather than an outcome of its own. All of the literature on respondents characteristics and GBD disability weights methodology focused on paired comparison questions[16–18]. There has been however some exploration on whether sex and other demographic variables affect the valuation of health with SG and TTO methodologies. In their review of the literature on TTO, Arnesen et al. 2005, found that in the few studies that assessed demographic differences in health valuation, there was evidence of women having lower health valuations than men, and that as age increased, valuations would decrease as well[6]. TTO weights range from 0 to 1, where 1 represent death and 0 perfect health, so lower valuations of health would mean women are less likely to trade years of healthy life to live a longer life in disability. In TTO studies with bigger sample sizes the sex pattern was mixed. In a UK study, Dolan et al. 2002 found that women had lower valuations than men and that these results were driven by big differences in severe health states[19]. In another UK sample Kharroubi et al. 2007 found health states to be higher for females[20]. While Zhuo et al. 2018 found no significant difference by sex in a Chinese sample representative of its population[21].

In the previous chapter we found that differences by sex on the valuation of PHE questions drive disability weights to vary greatly. However, we know little about how respondent characteristics like sex, age and education affect health valuation when using PHE questions. Although there is some literature on trade-off methods that discussed these issues, the little evidence that exists focuses on SG and TTO, which are methods used mostly for the valuation of health or to derive summary measures of population health in high income countries. The World Health Organization's Choosing Interventions that are Cost Effective (CHOICE) guidelines recommend the use of DALYs for low and middle income countries[22]. As such, differences in health valuation by sex in the PHE technique is not only valuable to understand for the estimation of disability weights for the GBD, but it also addresses a gap in the literature of health valuation methods mostly used in low- and middle-income countries.

Purpose of study

The purpose of this study is to examine differences by sex of the respondents on the valuation of health using population health equivalence data from the GBD study.

Methods

Data

From 2013 onwards, each round of the GBD study uses the disability weights estimated in the GBD 2013 disability weights study. This study combined paired comparison responses from GBD 2010 and the European DW study but only uses population health equivalence data from the 2010 study. Although PHE data was collected in the European study it was discarded from disability weights estimation because of quality concerns[23]. Therefore, this study focuses only on GBD 2010 DW study PHE data as it is the sole PHE source used to estimate Years of Life Lived with Disability (YLDs) and Disability Adjusted Life Years (DALYs) in the GBD study. A total of 33,296 PHE questions were collected in 2010 and 2011 from a web survey of 11,129 respondents across 167 countries[3].

Outcome

In population health equivalence questions, respondents are asked to choose between two health programs that they think had greater overall population health: "The first program prevented 1,000 people from getting an illness that causes rapid death. The second program prevented a number of people from getting an illness that is not fatal but causes a lifelong health problem". The bids or

number of people in the second program varied randomly from a set of prespecified options that ranged from 1,100 to 10,000 people. Similarly, the lifelong health problem that the second program is averting varies randomly from a prespecified set of 30 health states. These 30 health states were chosen to represent a spectrum of severity. The ultimate aim was to use these responses to anchor paired comparison preferences in order to estimate disability weights on a 0 to 1 scale[3].

In this chapter, the outcome of interest is the female to male ratio in the odds of choosing the second program over the first program. In other words, the ratio compares females to males in terms of how likely they are to choose a health program that averts a health state over saving the lives of 1,000 individuals.

Statistical analysis

We used a marginal logistic regression, estimated through generalized estimating equation (GEE)[24], to model respondents' decisions on choosing the second program on the PHE data with the following specification:

$$logit\ \Pr(Y_{i,phe} | X_{i,phe}) = \beta_0 + X_{i,phe}\beta \qquad (1)$$

Where $Y_{i,phe}$ is a binary variable indicating whether the respondent chose the second program, $X_{i,phe}$ is a vector of covariates including the respondent's sex and education as indicator variables, and age as categorical variables. As described in chapter 2, almost all PHE respondents had more than 12 years of education. In this analysis education was coded as an indicator variable with either more than 13 years education or 13 years or less, given the low frequency of respondents with no education or less than 12 years of education. We also included indicator variables for the health state and bids in the PHE question. Our GEE used an exchangeable working correlation structure and Huber-White robust standard errors[25] to account for within respondent correlation. We estimated the expected value of choosing the second program for both females and males conditional on the covariates described. With this information we estimated the female to male ratio across 1,000 bootstrap iterations of our model coefficients and their respective 95% confidence intervals.

Sensitivity analysis

On average, respondents were asked three PHE questions, resulting in clustered responses within individuals. Our main model accounts for within-respondent correlation. However, we further assessed whether we underestimated the standard error of the covariates in the model by running

an alternative model. We ran a liner mixed-effects model (LMM)[24] where respondent id is specified as a random intercept. Although it would be preferable to run a generalized linear mixed-effects model (GLMM), we ran a LMM given the large number of observations and clusters in our data that make it computationally prohibitive to run a (GLMM). Again, the systematic component of our LMM would follow the same covariates as in equation 1, but it would account for the within cluster correlation by respondent through b_{0i} as specified in the equation below:

$$E(Y_{i,phe} | X_{i,phe}, b_{0i}) = \beta_0 + X_i\beta + b_{0i}$$

Where the cluster effects are assumed to follow a normal distribution, $b_{0i} \sim N(0, \sigma_b^2)$.

Finally, we tested whether PHE health valuation varies by sex of the respondents when we incorporate PHE data collected after the GBD 2010 study. We added study as a categorical variable and interacted sex with study to estimate the differential effects of sex by study. Two studies are added to the analysis, the European DW study from 2013 and the Japanese DW study data from 2019[23,26]. Both studies followed the GBD methodology to collect PC and PHE data, but used GBD 2010 DW study PHE data for their disability weights estimation due to quality concerns. Although the added PHE data was subject to high levels of measurement error, appendix 1 shows a discernible pattern by sex of the respondent that could be further explored. Appendix 2 contains a detailed description of both the European and Japanese data.

Results

In the GBD 2010 internet survey, there were 11,129 participants that responded to three PHE questions. Notably, 70 percent of the respondents were female and 94 percent of them were highly educated with more than 13 years of education. Age was mostly equally distributed across respondents, 31 percent of them were between 18 and 29 years old, 23 percent were between 30 and 39 years old, 17 percent were between 40 and 49 years old, and 29 percent were 50 years old or older. The distribution of respondents was skewed towards high income countries, with 85 percent of the respondents being from High Income countries mostly from the United States, Australia and Western Europe.

Figure 1 details the female to male ratio of the probability of choosing the second program as generating greater population health benefits across health states and bids; the y axis shows the female to male ratio, while health states are ordered from least to most severe and bids from smallest

to largest along the x axis. Table 1 presents the health states in Figure 1 along with the percentage of bids per questions in which the female to male ratio of the probability of choosing the program that averts illness in the population is greater than one. On average, females are more likely to choose the program that averts illness over death regardless of the bid and health state. Only for three health states "Hearing loss: moderate", "Distance vision moderate impairment", and "Spinal cord lesion at neck level (treated)", the ratio of females over males is less than one for the majority of the bids in each health state. Two of these health states are at the low end of the severity distribution, while "Spinal cord lesion at neck level (treated)" is the most severe of the 30 PHE health states surveyed.

Figure 1. Response probabilities in population health equivalence questions in GBD 2010 DW study

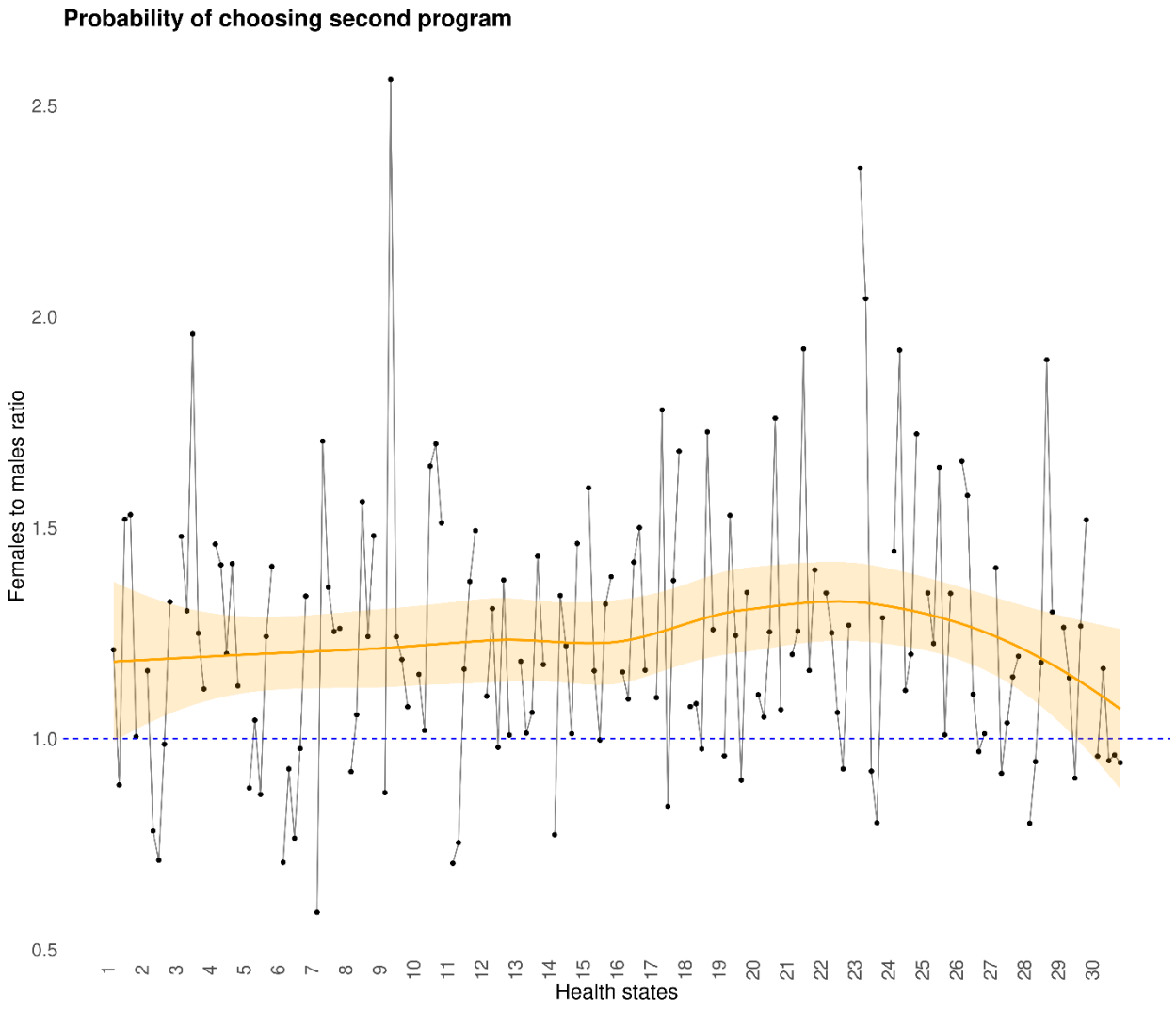


Table 1. Population health equivalence health states ordered by severity, percentage of bids with female to male ratios above 1

Health state	Severity	Percentage
Asthma, controlled	1	80
Hearing loss: moderate	2	40
Asthma, partially controlled	3	100
COPD and other chronic respiratory problems, mild	4	100
Itching and pain	5	60
Distance vision, moderate impairment	6	20
Musculoskeletal problems, lower limbs, mild	7	80
Angina pectoris, moderate	8	80
Hearing loss: complete	9	80
Diabetic neuropathy	10	100
Dementia, mild	11	60
Anemia, moderate	12	80
Anxiety disorders, moderate	13	100
Amputation of finger(s), excluding thumb: long term, with treatment	14	80
Acute myocardial infarction, days 3-28	15	80
Stroke, long-term consequences, moderate	16	100
Traumatic brain injury, long-term, moderate (with or without treatment)	17	80
Motor impairment, moderate	18	80
Cancer, diagnosis and primary therapy	19	60
Major depressive disorder, moderate episode	20	100
Decompensated cirrhosis of the liver	21	100
Distance vision, severe impairment	22	80
Parkinson's disease, moderate	23	60
Anxiety disorders, severe	24	100
COPD and other chronic respiratory problems, severe	25	100
Dementia, moderate	26	80
Distance vision blindness	27	80
Multiple sclerosis, severe	28	60
Stroke, long-term consequences, severe plus cognition problems	29	80
Spinal cord lesion at neck level (treated)	30	20

The differences between females and males can be formally observed in Table 2. Females had 1.34 (95% CI 1.25-1.43) times the odds of choosing the second program that averts disease as generating greater population health in comparison to males, keeping age, education, the health state and the bids constant. Interestingly we find that age and education do not meaningfully affect the probability of choosing the second program. These results are evident in Figure 2, where the ratio of females to males of the probability of choosing the program that averts illness is plotted

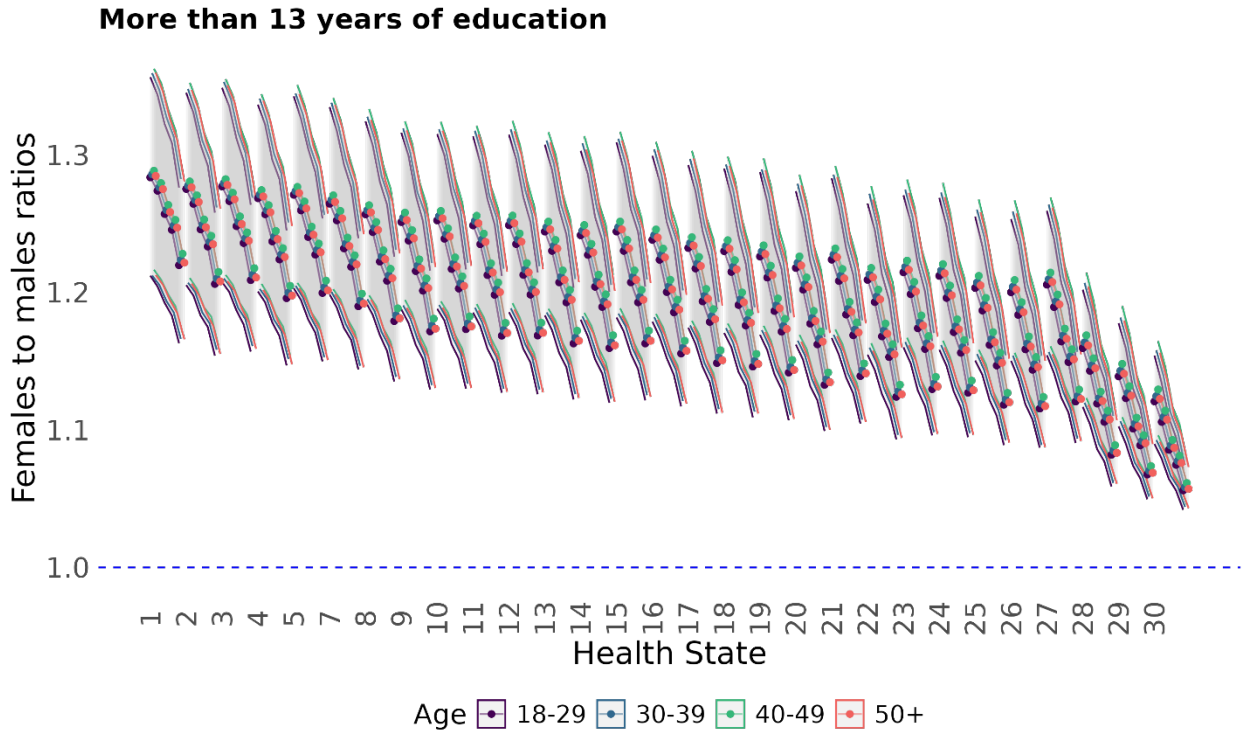
with bootstrapped 95 % confidence intervals. A clear sex pattern emerges, to increase overall population health females were more likely to choose the program that averts illness over immediate death regardless of age. These sex differences somewhat decrease as the bids become higher and the health states are more severe.

Table 2. Marginal logistic regression on choosing health program that averts illness

	OR (95%CI)	P values
Intercept	0.17 (0.14-0.2)	0
Female	1.34 (1.25-1.43)	0
Age		
30-39	0.95 (0.88-1.03)	0.24
40-49	0.9 (0.82-0.98)	0.02
50+	0.98 (0.9-1.05)	0.54
Education		
13y or less	1.05 (0.93-1.19)	0.43
N	33,170	
Respondents	11,087	
Working correlation	0.32	

With health state and bid fixed effects

Figure 2. Estimated ratio of probability of choosing second program, GBD 2010 PHE data



Sensitivity analysis

The primary results were robust to various functional forms of the regression model. Although in a different functional form, the LMM results in Appendix 3 helped us corroborate the meaningful and statistical difference between females and males. Appendix 1 shows that the pattern between sex of the respondent persists when we pool together data from the GBD 2010, the European and the Japanese disability weights studies: Females were more likely to choose the second program over the first. However, when we estimated the odds ratio of females in comparison to males per study, we observed differences by study. In the European and Japanese studies females had 1.15 (95% CI 1.11-1.19), and 1.24 (95% CI 1.2-1.28) times the odds, respectively, of choosing the program that averts disease over death compared to males, when keeping age, education, heath state and bid constant (Appendix 4). Although the difference in health valuation per sex persisted in different studies, they were less pronounced for Europe and Japan (Appendix 5).

Discussion

Females and males do not systematically differ in the ranking of preferences for health states, but they do disagree on their willingness to accept loss of life as a health program evaluator. Even when we take age and education into account, females are more likely to choose programs that avoid deteriorating health over preventing death for the relative few. These findings are robust to different model regressions and added disability weights data from more current studies. In other words, our PHE analysis show that females have a preference of quality of live over quantity of life.

The last two chapters have been devoted to describing in detail the sex differences in health valuation of GBD Disability weights and identifying where in the disability weight methodology the differences are originating. Although we have not examined why the differences between females and males exist, we have a couple of hypotheses that could help explain them. First, globally and on average men live shorter lives and women live longer lives with disabilities. In their 2024 study, Patwardhan et al. used the GBD study data from 2021 to compare morbidity and mortality differences between females and males. They find that males face higher DALYs than females and that the top health causes that lead to this difference is higher premature mortality due to cardiovascular diseases, cancers, road injuries and COVID-19. Females, on the other hand, live longer lives but have higher YLDs across all age groups in the life course, with the top contributors of morbidity being low back pain, depressive disorders, and headache disorders[27]. Given the sex disaggregated features of

disease burden, our results might be explained by females and males' different experiences with morbidity and mortality. Moreover, females may be more aware of the burden of chronic illness than males given their greater experience with providing care. Even if they do not experience illness firsthand, women are more exposed and responsible to aid other experiencing disease. Globally, women are disproportionally more likely to provide care – both formally and informally - for individuals with disabilities, chronic illness or the elderly[28,29]. Together, women's closer experience with disability in their own skin or as caregivers could explain divergent views between men and women on what type of program generate the greatest overall health. Further research could help explain these results, particularly using qualitative methods to probe small samples of PHE respondents to provide insights on the reasoning behind their decisions.

Before this analysis, literature on disability weights following the GBD 2010 methodology studied how differences in respondent characteristics affected health preferences through the analysis of paired comparison questions. As described in chapter two, the correlation between PC responses of females and males is high, even across different studies[17,18]. It is therefore not surprising that there has been little to no debate on health valuation discrepancies by sex of the respondent. PHE questions are mainly viewed as a methodological step in the anchoring of disability weights from 0 to 1, and have not been subject to stratified analysis based on characteristics of the respondents as PC data has. By diving deeper into PHE questions, our findings bring back to the frontline the old question in the health economics field of "whose preference count?"[30]. After much debate on whether experts, people affected by illness, or the general population should be valuating health, Salmon 2010 et al., and therefore the GBD study, decided that disability weights should be estimated from population surveys given that they are used to measure and advance population health goals[3]. Under this principal, the findings in this paper raise a logical concern: If females and males have systematic differences in their valuation of health, then a skewed PHE sample towards female would result in GBD disability weights that are closer to female preferences than males. Under the current methodology each respondent is weighted equally, however 70 percent of the PHE responses are female when roughly 50 percent of the global population is. It might be time to consider weighing PHE respondents to represent the sex composition of the global population, or alternatively be comfortable with knowing that in the GBD the answer to "whose preference count"? might be leaning a bit towards females.

Any discussion with such meaningful consequences in the global burden of disease should recognize the limitations of its analysis. Like other trade off health valuation methods, PHE questions are cognitively demanding[31]. From the results in different disability weights studies in Europe, Japan and China, researchers have hypothesized that to get a clear signal in the preferences of health valuation, respondents need to be highly educated[17,23,26]. The GBD DW study from 2010 only asked PHE questions in the web survey. Respondents were recruited through news items and editorials in scientific journals, announcements at scientific meetings and other forms of contact with global health professionals[3]. Moreover, people with access to internet services in 2010 would have been more educated than those without it. As a result, 94 percent of the respondents of the GBD DW 2010 study have more than 13 years of education, which meant our conclusions about how sex affects health valuation lacks variation in the education covariate. We address this issue somewhat with our sensitivity analysis. Both the European and Japanese studies have greater variability in the education of their respondents (Appendix 2), and we found no significant or meaningful association between education and the probability of choosing the second PHE program (Appendix 4). However, as substantially documented in the literature from these studies, the quality of the PHE data added in our sensitivity analysis is low. Further research on the relationship between education and sex of the respondents with high quality PHE data would be needed to fully address this limitation. Similarly, given the convenience sampling of PHE respondents, respondents from low- and middle-income countries were in the minority in our analysis. This is particularly relevant given that recent studies following the GBD DW methodology have found meaningful differences in estimated disability weights by regional context, even when newer data is collected in high income countries. Nomura et al. 2021 found their estimated Japanese DW to be two to three times larger than GBD 2013 disability weights for 20 of the health states valuated, and two to three times smaller than GBD 2013 disability weights for 23 of the health states valuated[26]. In China, Liu et al. 2022 found considerably lower disability weights for mental disorders, alcohol use and dementia than those in GBD 2013[17]. More data collection and PHE studies are needed across a variety of regions, particularly LMIC countries to test whether our sex pattern holds across cultural contexts. Beyond the improvements that greater variety of data would bring to the analysis of this research question, it is important to remember one of the goals of estimating DALYs is to inform policy makers of where to invest to improve health. It is in low-income settings where these tools are mostly needed, where budgets are already constrained. We should continue exploring whether our findings diverge for different regions of the world when a wider range of cultural and geographical contexts are included in the analysis.

Both limitations described seem to be inherent to trade-off methodologies, which are too cognitively demanding for their use in population surveys[31]. Of the three studies that collected PHE data following the GBD disability weight methodology after the 2010 study, only one used the data to estimate disability weights. Contrary to the European and Japanese DW studies, in the Chinese DW study of 2020, Liu et al. deemed the use of their collected PHE data appropriate to derive their study specific disability weights. This was not done without some logistical pain. PHE data from the Chinese study were deemed to be of good quality only after multiple levels of screening and a second round of data collection, where respondents were selected based on stricter criteria to improve PHE quality[17]. In other words, the lack of educational and geographic variation in PHE data is not only due to a lack of data collection efforts but also a consequence of demanding thought experiments from respondents. Innovation is needed to test different methods to collect high quality PHE data through population surveys. Further research could benefit by experimenting with alternative survey methodologies. Large Language Models, for example, could potentially be trained as interviewers, programming follow up clarifying questions if overall responses by respondents are not consistent across severity health state levels. Alternatively, practice trade-off questions similar to PHE questions but on different topics could be asked beforehand to train respondents to think in a trade-off mind set. In summary, our research highlights the need to innovate in the survey processes in search for higher signal-to-noise ratio of PHE responses for disability weights studies.

This study is an important contribution to the field. It is the first to study the demographic factors associated with respondent's valuation of health using population health equivalence questions. Previous literature has viewed PHE health valuation results mainly as a methodological step to anchor the preferences of PC health states between 0 and 1. In chapter two, we find that differences in health valuation through PHE questions translate into meaningful and significant differences in disability weights. In this chapter, our sex stratified PHE analysis shed light on a clear health valuation sex pattern. Females are more likely to prefer programs that avert disability as generating greater population health, or in other words, males are less willing to accept loss of life in comparison to disability. This new insight opens the discussion on how to make disability weights more representative of the global population. At minimum, it requires an evaluation of the common values used to estimate disability weights in the Global Burden of Disease, and its impacts on YLDs and DALYs.

Bibliography

1 Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. The purpose and scope of this book. In: Brazier J, Ratcliffe J, Saloman J, Tsuchiya A, eds. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press, 2016: 0.

2 Nord E. The Person-trade-off Approach to Valuing Health Care Programs. *Med Decis Making* 1995; **15**: 201–8.

3 Salomon JA, Vos T, Hogan DR, *et al*. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *The Lancet* 2012; **380**: 2129–43.

4 Essink-Bot ML, Bonsel GJ. How to derive disability weights. In: Summary measures of population health : concepts, ethics, measurement and applications. World Health Organization, 2002 https://apps.who.int/iris/handle/10665/42439 (accessed April 25, 2023).

5 Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Valuing health. In: Brazier J, Ratcliffe J, Saloman J, Tsuchiya A, eds. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press, 2016: 0.

6 Arnesen T, Trommald M. Are QALYs based on time trade-off comparable? – A systematic review of TTO methodologies. *Health Economics* 2005; **14**: 39–53.

7 Mansley EC, Elbasha EH. Preferences and person trade-offs: forcing consistency or inconsistency in health-related quality of life measures? *Health Economics* 2003; **12**: 187–98.

8 Burström K, Johannesson M, Diderichsen F. A comparison of individual and social time trade-off values for health states in the general population. *Health Policy* 2006; **76**: 359–70.

9 Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics* 2002; **11**: 447–56.

10 Drummond MF, Sculpher MJ, Torrance GW, Stoddart GL, J O. Methods for the Economic Evaluation of Health Care Programmes, 3rd edn. Oxford University Press, 2005 DOI:10.1093/oso/9780198529446.001.0001.

11 Lipman SA, Brouwer WBF, Attema AE. What is it going to be, TTO or SG? A direct test of the validity of health state valuation. *Health Economics* 2020; **29**: 1475–81.

12 Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goñi JM. Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health* 2019; **22**: 23–30.

13 Green C. On the societal value of health care: what do we know about the person trade-off technique? *Health Economics* 2001; **10**: 233–43.

14 Haagsma JA, Polinder S, Cassini A, Colzani E, Havelaar AH. Review of disability weight studies: comparison of methodological choices and values. *Population Health Metrics* 2014; **12**. DOI:10.1186/s12963-014-0020-2.

15 Charalampous P, Polinder S, Wothge J, von der Lippe E, Haagsma JA. A systematic literature review of disability weights measurement studies: evolution of methodological choices. *Archives of Public Health* 2022; **80**: 91.

16 Haagsma JA, Charalampous P. Deriving disability weights for the Netherlands: findings from the Dutch disability weights measurement study. *Popul Health Metrics* 2024; **22**: 26.

17 Liu X, Wang F, Zhou M, *et al*. Eliciting national and subnational sets of disability weights in mainland China: Findings from the Chinese disability weight measurement study. *The Lancet Regional Health – Western Pacific* 2022; **26**. DOI:10.1016/j.lanwpc.2022.100520.

18 Maertens de Noordhout C, Devleesschauwer B, Salomon JA, *et al*. Disability weights for infectious diseases in four European countries: comparison between countries and across respondent characteristics. *Eur J Public Health* 2018; **28**: 124–33.

19 Dolan P, Roberts J. To what extent can we explain time trade-off values from other information about respondents? *Social Science & Medicine* 2002; **54**: 919–29.

20 Kharroubi S, Brazier JE, O'Hagan A. Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. *Social Science & Medicine* 2007; **64**: 1242–52.

21 Zhuo L, Xu L, Ye J, *et al*. Time Trade-Off Value Set for EQ-5D-3L Based on a Nationally Representative Chinese Population Survey. *Value in Health* 2018; **21**: 1330–7.

22 Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health: an international perspective. In: Brazier J, Ratcliffe J, Saloman J, Tsuchiya A, eds. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford University Press, 2016: 0.

23 Haagsma JA, Noordhout CM de, Polinder S, *et al*. Assessing disability weights based on the responses of 30,660 people from four European countries. *Population Health Metrics* 2015; **13**. DOI:10.1186/s12963-015-0042-4.

24 Van Belle G, Lloyd DF, Heagerty PJ, Lumley T. Longitudinal Data Analysis. In: Biostatistics. John Wiley & Sons, Ltd, 2004: 728–65.

25 Van Belle G, Lloyd DF, Heagerty PJ, Lumley T. Association and Prediction: Linear Models with One Predictor Variable. In: Biostatistics. John Wiley & Sons, Ltd, 2004: 291–356.

26 Nomura S, Yamamoto Y, Yoneoka D, *et al*. How do Japanese rate the severity of different diseases and injuries?—an assessment of disability weights for 231 health states by 37,318 Japanese respondents. *Popul Health Metrics* 2021; **19**: 21.

27 Patwardhan V, Gil GF, Arrieta A, *et al*. Differences across the lifespan between females and males in the top 20 causes of disease burden globally: a systematic analysis of the Global Burden of Disease Study 2021. *The Lancet Public Health* 2024; **9**: e282–94.

28 Social protection Committee and the European Commission. Long-term care report. Trends, challenges and opportunities in an ageing society. 2021. https://www.ifsw.org/wp-content/uploads/2021/07/KE-09-21-202-EN-N-1.pdf (accessed April 26, 2025).

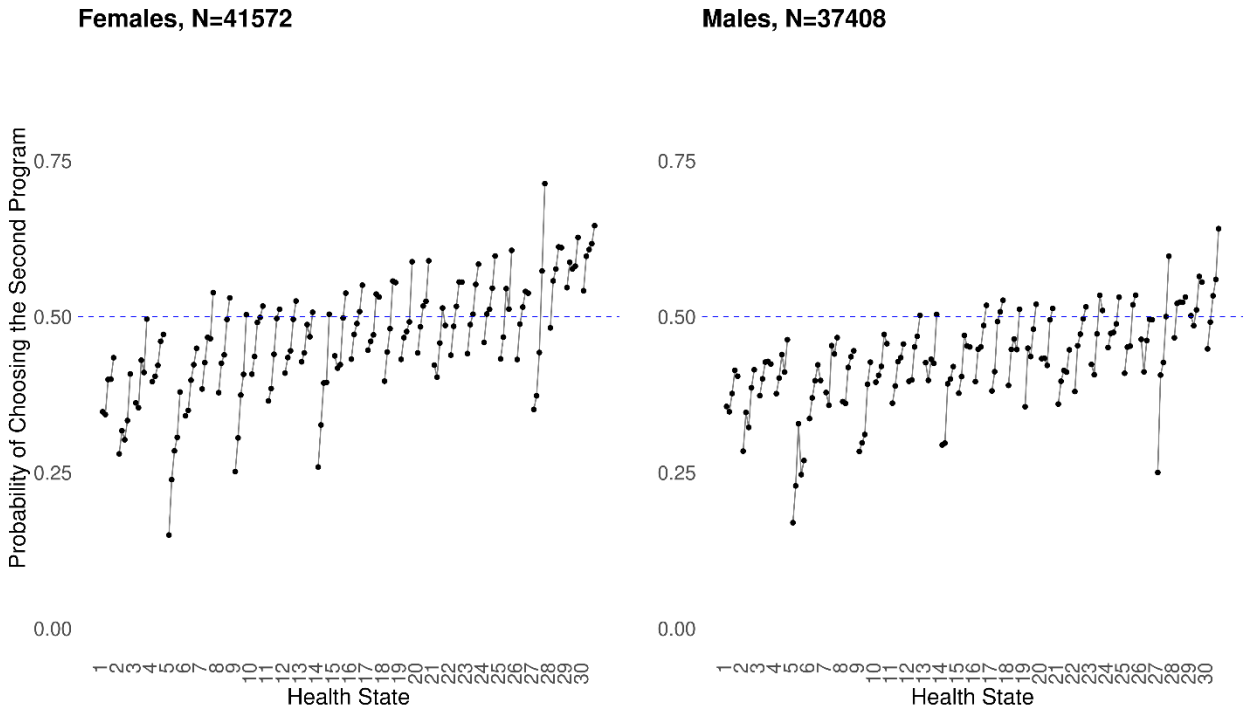
29 Skinner MS, Sogstad M. Social and Gender Differences in Informal Caregiving for Sick, Disabled, or Elderly Persons: A Cross-Sectional Study. *SAGE Open Nursing* 2022; **8**: 23779608221130585.

30 Dolan P. Whose preferences count? *Med Decis Making* 1999; **19**: 482–6.

31 Murray CJL, Salomon JA, Mathers CD, Lopez AD. Summary measures of population health: Conclusions and recommendations. In: Summary measures of population health : concepts, ethics, measurement and applications. World Health Organization, 2002 https://iris.who.int/handle/10665/42439 (accessed June 8, 2023).

Appendix

Appendix 1. Response probabilities in population health equivalence questions for GBD 2010, European and Japan DW studies



Appendix 2. Other disability weights study data

To study whether the differences in sex remain by different regions of the world we use a larger set of PHE data including the European and Japanese studies. As described in the previous chapter, current DW use population health equivalence questions collected in 2010 and 2011 from a web survey with 11,129 respondents[3]. For the European study, data were collected in Hungary, Italy, the Netherlands, and Sweden, chosen to represent Eastern, Southern, Central and Northern Europe. In Japan, Nomura et al. collected data on 37,318 Japanese respondents. Both the European and the Japanese DW studies selected respondents from internet panel lists based on certain demographic characteristics. The European study selected members 18 to 65 years old to match each country's percentage of age, sex and education level. Nomura et al. selected respondents from 18 to 70 years old to match Japanese 2015 National Census on ratios of age, gender and prefecture[26].

Across the European and Japanese studies sex is evenly distributed. Participants in the Japanese studies had more years of education than in the European study with only 4 percent of them with less than 12 years of education. In Italy, Hungary and the Netherlands more than a third of the participants had less than 12 years of education.

Although all three studies aimed to follow the GBD DW methodology, there were some differences in the PHE data that are worth mentioning. The first difference is in the number of bids. In GBD 2010 the bids were 1100, 1500, 2000, 3000, 5000 and 10000. In the European and Japanese DW studies the bids were 1500, 2000, 3000, 5000 and 10000. The second difference is on the health states chosen. In the European and Japanese studies itching or pain, and distance vision blindness were omitted. In the Japanese study, four of the remaining 28 health states used modified lay descriptions incorporated in the European study instead of the original 2010 lay descriptions for hearing loss moderate, hearing loss complete, spinal cord lesion at neck level (treated), and amputation of fingers(s) excluding thumb. Given how sensitive disability weights are to the lay descriptions presented to respondents, our analysis of pooled PHE data used the original 30 health states. In other words, all the comparisons of programs that include itching or pain, for example, would only have data from 2010 study because this health state was not included in the European and Japanese studies.

Appendix 3. OLS with individual random intercepts

	Coefficient (95%CI)	P values
Intercept	0.13 (0.04-0.22)	0.01
Female	0.06 (0.05-0.07)	0
Age		
30-39	-0.01 (-0.03-0.01)	0.24
40-49	-0.02 (-0.04-0.00)	0.02
50+	0.00 (-0.02-0.01)	0.63
Education		
13y or less	0.01 (-0.01-0.04)	0.34
N	33,170	
Respondents	11,087	
ICC	0.32	

With health state and bid fixed effects

Appendix 4. Marginal logistic regression on choosing health program that averts illness of pooled PHE data

	OR (95%CI)	P values
Intercept	0.21 (0.19-0.24)	0
Female	1.29 (1.21-1.38)	0
Study		
Europe	1.2 (1.12-1.27)	0
Japan	1.98 (1.86-2.1)	0
Female*Europe	0.88 (0.84-0.93)	0
Female*Japan	0.95 (0.9-1.01)	0.1
Age		
30-39	0.98 (0.95-1.01)	0.22
40-49	0.98 (0.95-1.01)	0.22
50+	0.94 (0.91-0.97)	0
Education		
13y or less	0.98 (0.96-1.01)	0.2
N	221,168	
Respondents	78,960	
Working correlation	0.37	

With health state and bid fixed effects

Appendix 5. Estimated ratio of probability of choosing second program. GBD 2010, Europe and Japan PHE data