Title

Opening the Black Box: Estimation of Targeted Effects in Causal Mediation Analysis - With Applications to Global Health and Occupational Epidemiology

Author

Wang, Aolin

Publication Date

2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Opening the Black Box:

Estimation of Targeted Effects in Causal Mediation Analysis

With Applications to Global Health and Occupational Epidemiology

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Epidemiology

by

Aolin Wang

2016

ABSTRACT OF THE DISSERTATION

Opening the Black Box:

Estimation of Targeted Effects in Causal Mediation Analysis

With Applications to Global Health and Occupational Epidemiology

by

Aolin Wang

Doctor of Philosophy in Epidemiology

University of California, Los Angeles, 2016

Professor Onyebuchi Aniweta Arah, Chair

Recent work has considerably advanced the definition, identification and estimation of different types of effects in causal mediation analysis. It extended the traditional approaches for mediation analysis by using the counterfactual or potential-outcome framework to allow for nonlinearities and exposure-mediator interaction. Despite the various estimation methods and statistical routines being developed, a unified approach is lacking, which incorporates recently introduced causal decompositions. Also, relatively few studies explored scenarios with more than one mediator. In this work, we used causal diagrams and potential-outcome framework to contribute to the literature on causal mediation analysis. We first provided a unified framework for estimating targeted effect(s) from the most recent 4-way decomposition in the single-mediator

setting. We demonstrated that g-computation, implemented via Monte Carlo simulations in standard statistical software, can offer such unification and is flexible in accommodating different types of exposure, mediator and outcome variables. We also extended some of the existing estimation techniques to more complicated mediation settings that involve contextual exposure, intermediate confounding, multiple causally ordered mediators, time-varying mediators, and time-to-event outcomes. We applied regression-based techniques, g-computation, and inverse-probability-weighted (IPW) fitting of marginal structural model (MSM) to investigate mechanisms underlying the effects of human development on individual health, the health disparity in education, and the effect of different physical activity domains on acute myocardial infarction. The flexibility of g-computation comes at a large cost: it becomes computationally intensive as the number of variables and sample size grow. Alternatively, mediator and outcome regression-based methods and IPW fitting of MSM can be applied in general linear systems and survival context respectively. The use of causal inference techniques did not preclude the possibility of model misspecification and the presence of uncontrolled confounding, which may bias our results. Future work should explore the properties of different estimation techniques and their use in estimating targeted quantities, and incorporate sensitivity analysis for uncontrolled confounding in causal mediation analysis.

The dissertation of Aolin Wang is approved.

Abdelmonem A. Afifi

Haroutune K. Armenian

Niklas Krause

Onyebuchi Aniweta Arah, Committee Chair

University of California, Los Angeles

2016

DEDICATION

To my parents, Hui Wang and Jia Lin,

and to my beloved husband, Chun-Jun Guo.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF ABBREVIATIONS

AMI	Acute myocardial infarction
BMI	Body mass index
CDE	Controlled direct effect
CDE_{sto}	Stochastic controlled direct effect
CHD	Coronary heart disease
CI	Confidence interval
CVD	Cardiovascular disease
DAG	Directed acyclic graph
DE	Direct effect
EE	Energy expenditure
GDP/c	Gross Domestic Product per capita
HDI	Human development index
HDL-C	High density lipoprotein cholesterol
HR	Hazard ratio
IHD	Ischemic heart disease
IPW	Inverse-probability-weighted
LMICs	Low- and middle-income countries
LTPA	Leisure-time physical activity
MIE	Mediated interaction effect
MSM	Marginal structural model
NCDs	Non-communicable diseases

NDE	Natural direct effect
NIE	Natural indirect effect
OPA	Occupational physical activity
PAI	Portion attributable to interaction
PDE	Pure direct effect
PE	Portion eliminated
PIE	Pure indirect effect
RAS	Relative aerobic strain
RERI	Relative excess risk for interaction
RIE	Reference interaction effect
RMSE	Root mean square error
SD	Standard deviation
SEM	Structural equation model
SES	Socioeconomic status
SUTVA	Stable unit treatment value assumption
TDE	Total direct effect
TE	Total effect
TIE	Total indirect effect
WHO	World Health Organization
WHS	World Health Survey

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my mentor and the chair of my dissertation committee, Dr. Onyebuchi Arah, for his constant inspiration, encouragement, guidance and support. I would like to thank Dr. Arah for giving me the opportunity to work with him during summer 2010 and introducing me to the wonderland of epidemiologic methods and causal inference. Without his help and support, I could not even start, let alone finish, this rewarding doctoral journey at UCLA. Through these years, he has set an example of excellence as a researcher, mentor, instructor, and role model. It has been the greatest fortune in my education to be able to work with and learn from him.

I am also very grateful to Dr. Niklas Krause, the principal investigator for the Kuopio project and my dissertation committee member, for his mentorship and many constructive comments that he provided on the conception, analysis, and writing of my projects. My gratitude is extended to other committee members, Drs. Abdelmonem Afifi and Haroutune Armenian, for their expertise, encouragement, invaluable advice and strong support. Special thanks to Dr. Beate Ritz for encouraging and helping me to pursue my career goals.

I would also like to thank my dear friends, both in Los Angeles and in Guangzhou, for making life fun, and providing help and emotional support during challenging moments. I am grateful to have them to share bits and pieces in my life. I would like to acknowledge the wonderful teachers I had at UCLA, the staffs at the UCLA Department of Epidemiology for their administrative support, the UCLA Institute for Digital Research and Education for providing

helpful statistical training, and the Southern California NIOSH Education and Research Center for providing training in occupational safety and health.

Last but not least, my heartfelt appreciation goes to my parents, my first teachers who believed in me and always made education possible for me, and to Chun-Jun, my husband, my best friend, and the love of my life, who has always been there for me, rain or shine.

This dissertation was supported by a doctoral scholarship from the China Scholarship Council (CSC) and the Dissertation Year Fellowship from UCLA. I greatly appreciate and acknowledge the funding support for my doctoral training from the CSC, UCLA, the American Association of University Women, Fielding School of Public Health, the Department of Epidemiology, Dr. Raymond Goodman, and the Chinese-American Engineers and Scientists Association of Southern California.

Chapter 2 is incorporated from Wang, A., Arah, O. A. G-computation demonstration in causal mediation analysis. *Eur J Epidemiol*. 2015 Oct;30(10):1119-1127. Doi: 10.1007/s10654-015-0100-z.

VITA

EDUCATION

2013	Master of Science in Epidemiology University of California, Los Angeles, CA, USA
2011	Bachelor of Medicine Fudan University, Shanghai, China

RESEARCH EXPERIENCE

2013 – 2014	Research Assistant EU 7th Framework Program for Research: Deepening our Understanding of Quality Improvement in European Hospitals (DUQuE)
2011 – 2012	Research Assistant Center for Global and Immigrant Health University of California, Los Angeles, CA, USA
2010	Research Intern Cross-disciplinary Scholars in Science and Technology (CSST) Program Department of Epidemiology University of California, Los Angeles, CA, USA

PEER-REVIEWED PUBLICATIONS

Wang, A., Arah, O. A. G-computation demonstration in causal mediation analysis. *Eur J Epidemiol*. 2015 Oct;30(10):1119-1127.

Wang, A., Arah, O. A. Body Mass Index and Poor Self-Rated Health in 49 Low-Income and Middle-Income Countries, By Sex, 2002–2004. *Prev Chronic Dis*. 2015;12:150070.

Wang, A., Arah, O. A., Kauhanen, J., Krause, N. Work schedule and 11-year progression of carotid atherosclerosis in middle-aged Finnish men. *Am J Ind Med*. 2015 Jan;58(1):1-13.

Sunol R., Wagner C., Arah O. A., Kristensen S., Pfaff H., Klazinga N., Thompson C. A., **Wang A.**, DerSarkissian M., Bartels P., Michell P.; Groene O., on behalf of the DUQuE Project Consortium. Implementation of departmental quality strategies is positively associated with clinical practice: results of a multicenter study in 73 hospitals in 7 European countries. *PLoS One*. 2015 Nov;10(11):e0141157.

Wang, A., Stronks, K., Arah, O. A. Global educational disparities in the associations between body mass index and diabetes mellitus in 49 low-income and middle-income countries. *J Epidemiol Community Health.* 2014 Aug;68(8):705-711. **[Editor's choice]**

Groene, O., Sunol, R., Klazinga, N. S., **Wang, A.**, Dersarkissian, M., Thompson, C. A., Arah, O. A. Involvement of patients or their representatives in quality management functions in EU hospitals: implementation and impact on patient-centred care strategies. *Int J Qual Health Care.* 2014 Apr;26 Suppl 1:81-91.

Shaw, C. D., Groene, O., Botje, D., Sunol, R., Kutryba, B., Klazinga, N., Bruneau, C., Hammer, A., **Wang, A.**, Arah, O. A., Wagner, C. The effect of certification and accreditation on quality management in 4 clinical services in 73 European hospitals. *Int J Qual Health Care.* 2014 Apr;26 Suppl 1:100-107.

Xie, Z., **Wang, A.**, Chen, B. Nurse burnout and its association with occupational stress in a cross-sectional study in Shanghai. *J Adv Nurs.* 2011 Jul;67(7):1537-1546. **[Co-first author]**

ORAL PRESENTATIONS

Wang, A., Nianogo, R. A., Arah, O. A. G-Computation and Inverse Probability of Treatment Weighted Estimation of Average Treatment Effects among the Treated and the Untreated. *Society for Epidmeiologic Research Annual Meeting: Concurrent Contributed Session*. Denver, CO, June 2015.

Wang, A., Arah, O. A., Kauhanen, J., Krause, N. Weekend shift work and progression of atherosclerosis. *Fall Interdisciplinary Workshop by Southern California NIOSH Education and Research Center*. Los Angeles, CA, Sept 2014.

Wang, A., Arah, O. A., Kauhanen, J., Krause, N. Work Schedule and 11-year progression of carotid atherosclerosis in middle-aged Finnish men. *IEA World Congress of Epidemoilogy*: *Occupational Health*. Anchorage, AK, Aug 2014.

Wang, A., Arah, O. A. G-Computation of Natural Effects in Causal Mediation Analysis. *Society for Epidmeiologic Research Annual Meeting: Intersections between Traditional Epidmeiology Methods and Simulation Modeling*. Seattle, WA, June 2014.

Chapter 1. Introduction

1.1 Importance of Mediation Analysis

The traditional approach to chronic disease epidemiology has been labeled as ”black box” paradigm.[1] This approach largely focuses on establishing exposure-disease association but paid little attention to the mechanisms underlying such association. By comparison, we are able to see, via mediation analysis, a more comprehensive picture of why and how an exposure has the effect that it does on disease. Mediation analysis can inform strategic interventions to block the harmful effect(s) of exposure at different points along the causal chain.

By probing for mechanisms from a cause to its effect, mediation analysis helps improve causal inference. First, identification of a hypothesized mediator can provide evidence that the observed relationship between exposure and disease is causal.[1] For instance, it has been suggested that between 40% and 60% of the protective effect of alcohol is mediated through increased levels of high density lipoprotein cholesterol (HDL-C).[2,3] This finding supports the hypothesis of a protective effect of moderate alcohol consumption against coronary heart disease (CHD).

Second, in certain cases, the overall effect of an exposure seems to be weak or even non-exist due to the cancellation of effect via different mechanisms. In the alcohol-CHD example, the beneficial effect of alcohol via increasing HDL-C might be offset to some degree by the alcohol’s adverse effects on increasing blood pressure.[4] Mediation analysis provides a tool to uncover such mechanisms by quantifying mediated effects via different pathways.

Third, it allows us to test pathway-specific hypotheses. As the socioeconomic status (SES) has been consistently linked to various health outcomes, the interest of social scientists shifts to answering the mechanisms that explain the socioeconomic gradient in health.[1] For example, we can use mediation analysis to test if SES affects health through providing more resources for people to maintain good health.

Regarding policy implication, mediation analysis offers quantitative evidence for decision-making. If low neighborhood SES affects obesity incidence mainly through restricting people's food choice to fast food or processed food, priority might be given to affordable and accessible healthier food options in the efforts to fight obesity. Mediation analysis also helps us in evaluating and improving an intervention. If we know that a multifaceted intervention improves a certain health outcome, a more important question is which components of such intervention are most effective in achieving the desired outcome. This is crucial for future implementation and refinement of the intervention, especially when it is carried out in places with scant resources. By measuring mediators along the proposed pathways and quantifying each pathway, we are able to evaluate through which pathways the intervention has bigger impact. Accordingly, this intervention can be improved to increase its overall efficacy by focusing on the components that leads to those effective mediators.

Mediation analysis has its roots in social science and psychometrics, dating back to Wright's[5,6] method of path analysis. The adoption of this method was delayed in epidemiology, mainly because the strong assumptions it imposes, i.e., linearity and no exposure-mediator interaction, are less plausible with epidemiologic data. Recent work has considerably advanced the

definition, identification and estimation of different types of effect in mediation analysis defined under the potential outcome framework and incorporate non-linear models and interactions.[7–12] Next, we briefly reviewed the most commonly used method for assessing mediation and its limitations, and how causal mediation analysis developed under the potential outcome framework can overcome these limitations.

1.2 Mediation Analysis in the Parametric Tradition and Its Limitations

In social science, the most widely used method for assessing mediation is the four-step approach based on traditional linear structural equation models (SEMs).[13–15] Let us consider a basic mediation structure represented in Figure 1.1 with single exposure or treatment (*X*), single outcome (*Y*), and single mediator (*M*). We wish to assess the role of *M* in transmitting the effect of *X* on *Y*. Throughout this dissertation, we will assume that *X* occurs before *M*, which occurs before *Y*.

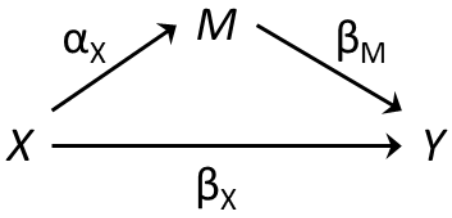

Figure 1.1 Graphical representation of a causal mechanism and the corresponding structural parameters involving three variables in a linear system.

In a linear system, the causal structure can be represented as linear structural equations:

$$x = \gamma_X + \varepsilon_X$$

$$m = \alpha_M + \alpha_X x + \varepsilon_M$$

$$y = \beta_Y + \beta_X x + \beta_M m + \varepsilon_Y$$

where lower-case symbols (x, m, y) represent the values that the variables (X, M, Y) may take, and the sum of the intercept and error term represents the variation explained in M, and Y by omitted factors U_X, U_M, U_Y. These omitted factors are unknown or unmeasured causes of X, M, Y and assumed to be mutually independent and thus are not drawn in Figure 1.1 above (i.e. a graph in which the direction of each edge represents causation and all the common parents of pairs of nodes are depicted). Structural parameters α_X, β_X, and β_M need to be estimated from the data. Then following the four steps: (1) estimate and test the total effect of X on Y, using a model without the mediator (defined as $\tau = \beta_X + \alpha_X\beta_M$), (2) estimate and test the X-M path, (3) estimate and test the M-Y path conditional on X, (4) estimate and test the X-Y path conditional on M (β_X), to see if it is zero, one can infer the presence of complete or partial mediation. Accordingly, indirect effect is defined as $\alpha_X\beta_M$.

In the health sciences, "difference method" is often used to estimate the indirect effect of X on Y through M. It considers the outcome model under two adjustment schemes: with and without adjusting for the mediator. Then it takes the difference in the coefficients for the exposure as the measure of indirect effect. Advantages include its simplicity and reliance on standard regression. However, it has two major drawbacks. First, it will lead to distorted and irreconcilable results when non-linearities or interactions are present in the underlying causal mechanism.[16–19] Second, its validity depends on the untested assumption that there is no uncontrolled confounding between X-Y, X-M, and M-Y relationship. Successful randomization of the exposure will support the assumption of no uncontrolled confounding of the X-Y and X-M relationship but will not guarantee the absence of uncontrolled confounding of the M-Y relationship.[20] If there is an unmeasured common cause L (Figure 1.2 (a), (b) in the next section) between M and Y, adjusting

for M will create spurious correlation[1] between X and Y and thus prevent the proper estimation of β_X or β_M.[21–23]

In the last two decades, several researchers have extended mediation analysis to non-linear structural equations system and relaxed the "no interaction between mediator and outcome" assumption,[7–12] thanks to the symbiosis of causal diagram and counterfactual thinking. Below we will define the common mediation parameters controlled direct effect (CDE), pure direct effect (PDE), and total indirect effect (TIE) using causal diagrams [24,25] and the potential outcomes (counterfactual) framework.[26]

A note on mediator and modifier (moderator)

Older literature frequently used "mediator" and "moderator" interchangeably until the clarification by Baron and Kenny.[13] "Mediator" refers to intersecting variable on the causal chain from the exposure to the outcome whereas "moderator", also termed "modifier" in epidemiologic studies, refers to any variable that interacts (either biologically or statistically) with another variable in affecting the outcome. A modifier can be a mediator or a covariate. Both mediator and modifier are relative to the exposure-outcome relation of interest. The presence of a

[1] An example of such phenomenon, called "collider bias" in epidemiology, is the sprinkler example. Assume X=1 means the sprinkler is on and L=1 means it is raining and the two events are independent. Both X=1 and L=1 lead to M=1 (arrows going from X and L to M), which denotes the floor is wet. Thus, by observing the floor is wet (meaning conditioning on a collider) and sprinkler is off, we can infer that it must be raining (thus creating dependency between X and L).

mediator causes the phenomenon "mediation". Similarly, the presence of a modifier causes the phenomenon "effect measure modification", or sometimes called "heterogeneity" or "(statistical) interaction". Throughout the dissertation, we will use "modifier" to be consistent with the epidemiologic literature. "Effect measure modification", "heterogeneity", and "(statistical) interaction" will be used interchangeably.

1.3 Mediation Analysis under the Potential Outcome Framework

Brief review of directed acyclic graph (DAG) in causal mediation analysis

Directed acyclic graphs (DAGs) are graphical representations of causal structures that contain no feedback loops (acyclic) (Figure 1.2). In a DAG, variables are called nodes and connectors (arrows) are called edges. The direction of each edge represents causation and all the common parents of pairs of nodes are depicted. We use a square bracket around a node to indicate conditioning on the variable. A ***collider*** is a variable on a specific path where two arrowheads meet (M is a collider on the path $X \rightarrow M \leftarrow L$ in Figure 1.2 (a)). A path is said to be open or unblocked unconditionally if there is no collider on the path. Otherwise, the path is closed or blocked. In Figure 1.2 (a), path $X \rightarrow M \leftarrow L$ is blocked by the collider M, whereas path $X \rightarrow M \rightarrow Y$ is unblocked. Conditioning on a non-collider M on a path blocks the path at M. However, conditioning on a collider M, or any consequence of M, or any combination of M or its consequence, opens the path at M. In Figure 1.2 (b), $X \rightarrow [M] \rightarrow Y$ is blocked at M but path $X \rightarrow [M] \leftarrow L$ is opened at M due to conditioning on M. A ***backdoor path*** between any two nodes is a non-causal path between them that starts with an arrow pointing into the starting node, e.g., $M \leftarrow L \rightarrow Y$ is backdoor path between M and Y. Thus, if we are interested in the causal effect of M on

Y, we need to block such backdoor path via conditioning on *L*, which is thus called a confounder of the effect of *M* on *Y*. DAGs have been discussed extensively in the literature.[24,25,27]

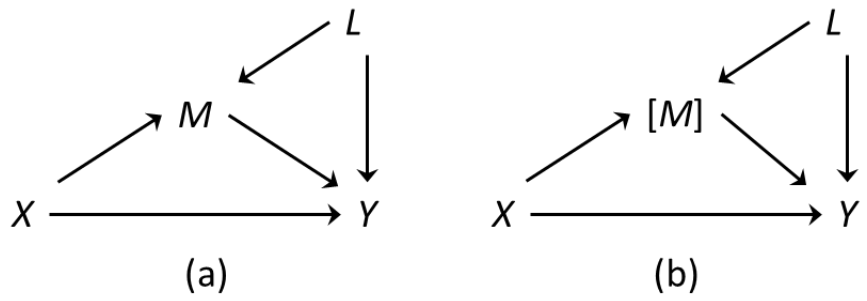

Figure 1.2 DAGs depicting a causal mechanism involving exposure *X*, mediator *M*, outcome *Y*, and a confounder *L* of the *M*-*Y* relationship without (a) and with (b) conditioning on *M*.

Notation: observed random variables and potential outcomes

Consider the same structure as presented in Figure 1.1. Let *X* denote the exposure of interest, *Y* the outcome of interest, *M* the mediator of interest, and *Z* a set of covariates not affected by the exposure but which are assumed to be sufficient for confounding control for total, direct and indirect effects estimation. Let Y_x and M_x denote respectively the potential values of the outcome and mediator that would have occurred had exposure *X* been set, possibly counter to fact, to a specific value *x*. Similarly, let Y_{xm} denote the potential value of *Y* that would have occurred had *X* and *M* been set, possibly counter to fact, to *x* and *m* respectively. We use $Y_{xM_{x^*}}$ to express potential outcome value had the exposure *X* been set to *x* and *M* to M_{x^*}. For simplicity, we will use 1 and 0 to represent index and reference values of exposure respectively.

The average total effect (TE), defined as $E[Y_1 - Y_0]$, compares exposure level 1 to 0, while allowing the mediator to obtain its natural value under each exposure level across the population. By assuming generalized consistency [10], TE can also be defined as $E[Y_{1M_1} - Y_{0M_0}]$.

The average controlled direct effect (CDE), defined as $E[Y_{1m} - Y_{0m}]$, compares exposure level 1 to 0 while fixing the mediator to a specific level *m*. The CDE estimates the effect of *X* on *Y* while fixing *M* to *m* for every individual in the population and it can be different for different levels of *m* in the presence of *X*-*M* interaction.[8,11]

Total effect can be decomposed into pure direct effect (PDE) and total indirect effect (TIE) as follows: $TE = E[Y_{1,M_0} - Y_{0,M_0}] + E[Y_{1,M_1} - Y_{0,M_1}] = E[Y_{1,M_0} - Y_0] + E[Y_1 - Y_{0,M_1}]$. The average PDE compares exposure level 1 to 0 while the mediator *M* is set to the natural value it would have attained under the reference level 0 of exposure (i.e., M_0). On the other hand, the average TIE compares mediator level M_1 to M_0 while setting exposure to index level 1.

We will use an example that has the same structure as presented in Figure 1.1 to illustrate how direct and indirect effects defined under the potential outcome framework generalizes that under linear structural equation models. Assume *X*, *M*, and *Y* are continuous variables. We can express each variable as a function of its determinants as follows:

$$m = \alpha_M + \alpha_X \cdot x + \varepsilon_M \quad (1)$$

$$y = \beta_Y + \beta_X \cdot x + \beta_M \cdot m + \beta_{XM} \cdot x \cdot m + \varepsilon_Y \quad (2)$$

$$\int E(Y|x,m,c)dF(m|x,c)$$

$$= \beta_Y + \beta_X \cdot x + \beta_M \cdot (\alpha_M + \alpha_X \cdot x + \varepsilon_M) + \beta_{XM} \cdot x \cdot (\alpha_M + \alpha_X \cdot x + \varepsilon_M)$$

$$= \beta_Y + \beta_M\alpha_M + \beta_X \cdot x + \beta_M\alpha_X \cdot x + \beta_{XM}\alpha_M \cdot x + (\beta_{XM}\alpha_X \cdot x) \cdot x \quad (3)$$

For binary exposure variable that takes values 1(index) and 0 (reference), PDE and TIE can be expressed in terms of regression coefficients from equation (1) and (2):

$$PDE = \beta_X + \boldsymbol{\beta_{XM}}\alpha_M$$

$$TIE = \beta_M\alpha_X + \boldsymbol{\beta_{XM}}\alpha_X$$

In the absence of exposure-mediation interaction (i.e., $\beta_{XM} = 0$), direct (β_X) and indirect ($\beta_M\alpha_X$) effects coincide under the potential outcome framework and using the traditional SEM. Direct and indirect effects defined in terms of path tracing in linear SEMs can be considered a special case of the pathway-specific effects defined under the potential outcome framework. We will use "causal mediation analysis" to refer to mediation analysis under potential outcome framework.

1.4 Gaps in the Literature

The past decade has seen the flourishing of identification criteria and estimation techniques in causal mediation analyses. These estimation techniques are designed for the single-mediator setting and only a few extends the method to settings with intermediate confounding (that is, exposure-induced confounding of the mediator-outcome relation),[28] or with multiple

mediators,[29–31] or with time-varying exposure and mediators.[32] Among these estimation techniques for a single-mediator setting, with some notable exceptions,[33–35] relatively few approaches can incorporate general types of mediator and outcome variables. Despite the existing statistical programs and procedures, relatively few applied studies have been published,[36,37] and even fewer applied studies that considered situations beyond the single-mediator setting.[38] Reasons may be that some of the methods impose strong assumptions that applied research may not support such as no intermediate confounding or multiple mediators not being causally related to each other after conditioning on exposure.

This dissertation embraces the current advancement and aims to extend the use of causal mediation analyses by offering a unified framework for estimating targeted effects of interest and particularly in probing mediation questions in the global health and occupational health contexts. We begin with reviewing effect definition, identification and estimation, and providing a unified framework for effect estimation in the single-mediator setting (Chapter 2). Then we explore different estimation techniques in more complex settings involving intermediate confounding (Chapter 3), multiple causally ordered mediators (Chapter 4), and time-varying mediators (Chapter 5), based on research questions in global health and occupational health. For mediation analysis applied in specific contexts (Chapters 3-5), we define the targeted effects based on specific research questions, describe the identification criteria and estimation techniques, present and discuss the results. We conclude with a general discussion and some implications for future research.

1.5 Specific Aims for This Dissertation

The specific aims and a brief statement of the objective of each chapter are listed below:

1) To formalize and demonstrate the use of g-computation in estimating different targeted effects in causal mediation analysis (Chapter 2). This chapter provides a brief overview of causal mediation analysis under the most recent 4-way decomposition in the single-mediator setting and illustrates the utility of (parametric) g-computation with a partially simulated data set.
2) To explore the pathways from human development to individual health, that are through individual education or weight status (Chapter 3). This chapter involves quantifying education path-specific effect and body mass index path-specific effect (a scenario with education being an intermediate confounder) of human development, a contextual factor, on individual health.
3) To investigate the mediating role of social factors and health behaviors in explaining health disparities in education (Chapter 4). We decompose the total effect of education on health into a portion involving compositional factors and a portion involving health behaviors only, accounting for the fact that health behaviors can be affected by these compositional factors even after conditioning on education.
4) To assess both the modifying and mediating roles of leisure-time physical activity in the effect of occupational physical activity on acute myocardial infarction (Chapter 5). This chapter examines the complex interplay between these two types of physical activity on cardiovascular health by implementing both interaction and mediation analysis, and explores the mediating role of both baseline and follow-up leisure-time physical activity.

Chapter 2. G-Computation Demonstration in Causal Mediation Analysis

2.1 Abstract

Recent work has considerably advanced the definition, identification and estimation of controlled direct, and natural direct and indirect effects in causal mediation analysis. Despite the various estimation methods and statistical routines being developed, a unified approach for effect estimation under different effect decomposition scenarios is still needed for epidemiologic research. G-computation offers such unification and has been used for total effect and joint controlled direct effect estimation settings, involving different types of exposure and outcome variables. In this study, we demonstrate the utility of parametric g-computation in estimating various components of the total effect, including (1) natural direct and indirect effects, (2) standard and stochastic controlled direct effects, and (3) reference and mediated interaction effects, using Monte Carlo simulations in standard statistical software. For each study subject, we estimated their nested potential outcomes corresponding to the (mediated) effects of an intervention on the exposure wherein the mediator was allowed to attain the value it would have under a possible counterfactual exposure intervention, under a pre-specified distribution of the mediator independent of any causes, or under a fixed controlled value. A final regression of the potential outcome on the exposure intervention variable was used to compute point estimates and bootstrap was used to obtain confidence intervals. Through contrasting different potential outcomes, this analytical framework provides an intuitive way of estimating effects under the recently introduced 3- and 4- way effect decomposition. This framework can be extended to complex multivariable and longitudinal mediation settings.

2.2 Introduction

Recent work has considerably advanced the definition, identification and estimation of different controlled or natural effects in causal mediation analysis. As a result, various estimation methods such as linear structural equation modeling, outcome and mediator regression-based methods, parametric g-computation, inverse-probability-weighted (IPW) fitting of marginal structural models (MSMs), and sequential g-estimation[39,40] have been applied to mediation settings. With some notable exceptions,[33–35] relatively few approaches can incorporate general types of mediator and outcome variables. The simulation-based approach introduced by Imai et al.[35] and the regression-based approach proposed by Valeri & VanderWeele[34] are increasingly popular approaches; the former may be unfamiliar to epidemiologists, and the latter requires different regressions or approximations for different mediator and outcome types. Meanwhile, g-computation,[39] as an alternative for computing marginal effects over IPW fitting of MSM,[41] holds promise to provide a unified framework for effect estimation in causal mediation analysis, especially given its capability in dealing with time-varying exposure and confounding.[32,33] However, a didactic demonstration of the application of g-computation in causal mediation analysis, reflecting recent decompositions for mediation and interaction, is lacking.

In this paper, we demonstrate the utility of (parametric) g-computation in estimating various components of the total effect, such as (1) natural direct and indirect effects, (2) standard and stochastic controlled direct effects, and (3) reference and mediated interactions, using standard statistical software. We focus on marginal effects (standardized over covariates). The current approach extends the previous work on g-computation demonstration for total effect[41] and the *gformula* package (mediation option) in Stata[33] by showing the actual steps in estimation and

incorporating estimation for various component effects under 2-, 3- and 4-way effect decomposition.[42,43] The paper is organized as follows: we will first review the definition and identification criteria for different effects in mediation context. Then we will review the g-computation algorithm and introduce steps for mediation analysis. After an illustrative example using a partially simulated data set, we will discuss the strengths and limitations of g-computation and its relation to other existing estimation procedures. Readers familiar with the background material on mediation analysis under the potential outcomes framework can go directly to g-computation steps section.

2.3 Methods

2.3.1 Notation and definitions

Let *X* denote the exposure of interest, *Y* the outcome of interest, *M* the mediator of interest, and *Z* a set of covariates not affected by the exposure but which are assumed to be sufficient for confounding control for total, direct and indirect effects estimation. Throughout, we assumed *X* preceded *M*, which preceded *Y*. Let Y_x and M_x denote respectively the potential values of the outcome and mediator that would have occurred had exposure *X* been set, possibly counter to fact, to a specific value *x*. Similarly, let Y_{xm} denote the potential value of *Y* that would have occurred had *X* and *M* been set, possibly counter to fact, to *x* and *m* respectively. We use $Y_{xM_{x^*}}$ to express potential outcome value had the exposure *X* been set to *x* and *M* to M_{x^*}. Let *x* (index) and x^* (reference) denote two values of the exposure we wish to compare, and *m* (index) and m^* (reference) denote two controlled values of the mediator we wish to compare.

Total effect (TE) compares exposure level x to x^*, while allowing the mediator to obtain its natural value under each exposure level. Thus, by assuming generalized consistency,[44] TE can also be defined as $\mathrm{E}[Y_{xM_x} - Y_{x^*M_{x^*}}]$. TE can be decomposed into different types of component effects. There are two-way, three-way, and four-way decompositions of the total effect as presented in Table 2.1.[8,11,42,43] The counterfactual definitions are listed in Table 2.2 (left column). The choice of effect decomposition should be guided by substantive research questions (Table 2.3).

There are four types of "direct" effects. The standard controlled direct effect (CDE) compares exposure level x to x^* while fixing the mediator to a specific level. The CDE estimates the effect of X on Y while fixing M to m for every individual in the population and it can be different for different levels of m.[8,11] The stochastic CDE ($\mathrm{CDE}_{\mathrm{sto}}$) compares exposure level x to x^* while randomizing the mediator to a pre-specified distribution M'. Accordingly, the stochastic CDE subsumes both the standard CDE and the stochastic mediation contrast[37] in that the standard CDE corresponds to M' being a constant m for the total population (i.e., full intervention) while the stochastic mediation contrast corresponds to M' being a constant for a subset of the population and being the observed distribution for the rest (i.e., partial intervention).

The average pure direct effect (PDE) compares exposure level x to x^* while the mediator M is set to the natural value it would have attained under the reference level x^* of exposure (i.e., M_{x^*}). Accordingly, the average total direct effect (TDE) differs from PDE in that the mediator M is set to the natural value it would have attained under the index level x of exposure (i.e., M_x). Direct

effects are relative to the mediator M of interest, that is, they are effects through alternative pathways other than through M.

The average pure indirect effect (PIE) compares mediator level M_x to M_{x*} while setting exposure to reference level x^*. The average total indirect effect (TIE) differs from the PIE in that the exposure is set to index level *x*.

Reference interaction effect (RIE) and mediated interaction effect (MIE) capture the effect of *X* on *Y* due to interaction only and the effect of *X* on *Y* due to both interaction and mediation respectively.[43] RIE and MIE are sometimes combined to reflect the effect of *X* on *Y* due to overall interaction and termed "portion attributable to interaction" (PAI).[43]

Table 2.1 Summary of effect decomposition in causal mediation analysis[a]

N-way decomposition			**Counterfactual definition (individual level)**
2-way	TE[b]	= PDE + TIE	$= (Y_{xM_{x^*}} - Y_{x^*M_{x^*}}) + (Y_{xM_x} - Y_{xM_{x^*}})$
		= TDE + PIE	$= (Y_{xM_x} - Y_{x^*M_x}) + (Y_{x^*M_x} - Y_{x^*M_{x^*}})$
		= CDE + PE	$= (Y_{xm} - Y_{x^*m}) + [(Y_x - Y_{x^*}) - (Y_{xm} - Y_{x^*m})]$
3-way	TE	= PDE + MIE + PIE	$= \left(Y_{xM_{x^*}} - Y_{x^*M_{x^*}}\right) + (Y_{xM_x} - Y_{xM_{x^*}} - Y_{x^*M_x} + Y_{x^*M_{x^*}}) + (Y_{x^*M_x} - Y_{x^*M_{x^*}})$
			$= \left(Y_{xM_{x^*}} - Y_{x^*M_{x^*}}\right) + (Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_x - M_{x^*}) + (Y_{x^*M_x} - Y_{x^*M_{x^*}})$
		= CDE + PAI + PIE	$= (Y_{xm} - Y_{x^*m}) + (Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_x) + (Y_{x^*M_x} - Y_{x^*M_{x^*}})$
4-way	TE	= CDE + RIE	$= (Y_{xm} - Y_{x^*m}) + (Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_{x^*})$
		+ MIE + PIE	$+(Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_x - M_{x^*}) + (Y_{x^*M_x} - Y_{x^*M_{x^*}})$

[a] Table adapted from VanderWeele (2014) Tables 5-6.[43]

[b] TE: total effect, PDE: pure direct effect, TIE: total indirect effect, TDE: total direct effect, PIE: pure indirect effect, CDE: controlled direct effect (standard), PE: portion eliminated, MIE: mediated interaction effect (referred to as "INT_{med}" by VanderWeele), PAI: portion attributable to interaction, RIE: reference interaction effect (referred to as "INT_{ref}" by VanderWeele).

Table 2.2 Definition and empirical analog of total effect and component effects[a]

Effect	Counterfactual definition	Empirical analog[b]
TE[c]	$\mathrm{E}[Y_x - Y_{x^*}]$[d]	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z})P(m\|x,\boldsymbol{z}) - E(Y\|x^*,m,\boldsymbol{z})P(m\|x^*,\boldsymbol{z})\}P(\boldsymbol{z})$ [e]
PDE	$\mathrm{E}[Y_{xM_{x^*}} - Y_{x^*M_{x^*}}]$	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z}) - (Y\|x^*,m,\boldsymbol{z})\}P(m\|x^*,\boldsymbol{z})P(\boldsymbol{z})$[f]
TIE	$\mathrm{E}[Y_{xM_x} - Y_{xM_{x^*}}]$	$\sum_z \sum_m E(Y\|x,m,\boldsymbol{z})\{P(m\|x,\boldsymbol{z}) - P(m\|x^*,\boldsymbol{z})\}P(\boldsymbol{z})$
TDE	$\mathrm{E}[Y_{xM_x} - Y_{x^*M_x}]$	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z}) - (Y\|x^*,m,\boldsymbol{z})\}P(m\|x,\boldsymbol{z})P(\boldsymbol{z})$
PIE	$\mathrm{E}[Y_{x^*M_x} - Y_{x^*M_{x^*}}]$	$\sum_z \sum_m E(Y\|x^*,m,\boldsymbol{z})\{P(m\|x,\boldsymbol{z}) - P(m\|x^*,\boldsymbol{z})\}P(\boldsymbol{z})$[f]
$\mathrm{CDE}_{\mathrm{M=m^*}}$	$\mathrm{E}[Y_{xm^*} - Y_{x^*m^*}]$	$\sum_z \{E(Y\|x,m^*,\boldsymbol{z}) - E(Y\|x^*,m^*,\boldsymbol{z})\}P(\boldsymbol{z})$
$\mathrm{CDE}_{\mathrm{sto}}$	$\mathrm{E}[Y_{xM'} - Y_{x^*M'}]$	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z}) - E(Y\|x^*,m,\boldsymbol{z})\}P(m')P(\boldsymbol{z})$
RIE	$\mathrm{E}[(Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_{x^*})]$	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z}) - E(Y\|x,m^*,\boldsymbol{z}) - E(Y\|x^*,m,\boldsymbol{z}) + E(Y\|x^*,m^*,\boldsymbol{z})\}P(m\|x^*,z)P(\boldsymbol{z})$
MIE	$\mathrm{E}[(Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_x - M_{x^*})]$	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z}) - E(Y\|x,m^*,\boldsymbol{z}) - E(Y\|x^*,m,\boldsymbol{z}) + E(Y\|x^*,m^*,\boldsymbol{z})\}\{P(m\|x,\boldsymbol{z}) - P(m\|x^*,\boldsymbol{z})\}P(\boldsymbol{z})$
PAI	$\mathrm{E}[(Y_{xm} - Y_{xm^*} - Y_{x^*m} + Y_{x^*m^*})(M_x)]$	$\sum_z \sum_m \{E(Y\|x,m,\boldsymbol{z}) - E(Y\|x,m^*,\boldsymbol{z}) - E(Y\|x^*,m,\boldsymbol{z}) + E(Y\|x^*,m^*,\boldsymbol{z})\}P(m\|x,z)P(\boldsymbol{z})$

[a] Y: outcome, X: exposure, M: mediator, Z: covariates; *x* and *m* represent the index values whereas x^* and m^* represent the reference values.

[b] Under the stable unit treatment value assumption, consistency, conditional exchangeability, positivity, different types of effect can be identified and estimated using the empirical analogs We use $E(Y|x,m,\boldsymbol{z})$ as a shorthand for $E(Y|X=x,M=m,\boldsymbol{Z}=\boldsymbol{z})$, and $P(m|x,\boldsymbol{z})$ as a shorthand for $P(M=m|X=x,\boldsymbol{Z}=\boldsymbol{z})$.

[c] TE: total effect, PDE: pure direct effect, TIE: total indirect effect, TDE: total direct effect, PIE: pure indirect effect, CDE: controlled direct effect (standard), $\mathrm{CDE}_{\mathrm{sto}}$: stochastic controlled direct effect, RIE: reference interaction effect (referred to as "$\mathrm{INT}_{\mathrm{ref}}$" by VanderWeele), MIE: mediated interaction effect (referred to as "$\mathrm{INT}_{\mathrm{med}}$" by VanderWeele), PAI: portion attributable to interaction.

[d] Effects are defined as risk differences here but other measures of effects are possible (risk ratio, odds ratio etc.).

[e] For continuous M and Z, summations are replaced by integrals and the probability functions by appropriate density functions

[f] These two expressions are known as the mediation formula.[45]

Table 2.3 Potential research questions related to various effects under effect decomposition.

Effect	Research question
TE[a]	Overall, to what extent does *X* cause *Y*?
PDE	In particular, to what extent does *X* cause *Y* via pathways other than through *M*?
TIE	In particular, to what extent does *X* cause *Y* via *M* (i.e. due to *X* affecting *M* and subsequently, *M* affecting *Y*) and the possible interaction between *X* and *M* in affecting *Y*? In other words, the effect of exposure that "would be prevented if the exposure did not cause the mediator" (i.e. the portion of the effect for which mediation is "necessary") [47,48].
TDE	In particular, to what extent does *X* cause *Y* via pathways other than through *M*, allowing *M* to boost up or tune down such effect at the same time?
PIE	In particular, to what extent does *X* cause *Y* via *M* only (i.e. due to *X* affecting *M* and subsequently, *M* affecting *Y*), not accounting for the possible interaction between *X* and *M*? In other words, the effect that the exposure would have had if "its only action were to cause the mediator" (i.e. the portion of the effect for which mediation is "sufficient") [47,48].
CDE	What would be the effect of *X* on *Y*, when fixing *M* at a specific value for everyone in the population?
CDE_{sto}	What would be the effect of *X* on *Y*, when allowing *M* to attain certain controlled distribution (via intervention) in the population?
RIE	What would be the effect of *X* on *Y* that is due to interaction between *X* and *M* only?
MIE	What would be the effect of *X* on *Y* that is due to both interaction between *X* and *M* and the fact that *X* causes *M*?
PAI	What would be the effect of *X* on *Y* that is due to interaction between *X* and *M*, regardless whether *X* causes *M*?

[a] TE: total effect, PDE: pure direct effect, TIE: total indirect effect, TDE: total direct effect, PIE: pure indirect effect, CDE: controlled direct effect (standard), CDE_{sto}: stochastic controlled direct effect, RIE: reference interaction effect (referred to as "INT_{ref}" by VanderWeele), MIE: mediated interaction effect (referred to as "INT_{med}" by VanderWeele), PAI: portion attributable to interaction.

2.3.2 Assumptions for identification and estimation

To identify and estimate the effect decomposition quantities, we invoke the stable unit treatment value assumption (SUTVA),[39,46] and assumptions of consistency,[47] conditional exchangeability (no-uncontrolled-confounding), and positivity.[48] The conditional exchangeability assumption for mediation analysis includes the following:[11,28] (i) the effect of the exposure X on the outcome Y is unconfounded conditional on a set $\boldsymbol{Z}$ of measured covariates; (ii) the effect of the mediator M on Y is unconfounded conditional on both X and $\boldsymbol{Z}$; (iii) the effect of X on M is unconfounded conditional on $\boldsymbol{Z}$. For identifying standard CDE, (i) and (ii) are sufficient but for stochastic CDE, all three assumptions are needed. Successful randomization of the exposure will support the assumption of no uncontrolled confounding of the exposure-mediator and exposure-outcome relations but will not guarantee the absence of uncontrolled confounding of the mediator-outcome relation.[20] In the presence of possible violation of this assumption, sensitivity analysis are needed.[35,49] To identify natural effects, a fourth conditional exchangeability assumption is needed: (iv) none of the mediator-outcome confounders are affected by exposure. Assumption (iv) is known as the cross-world independence assumption[50] because it requires that conditional on $\boldsymbol{Z}$, the mediator that would have been observed in a world under $X=x^*$ is independent of the outcome that would have been observed in a world under $X=x$ (i.e. $M_{x^*} \perp\!\!\!\perp Y_{xM_x}$). This assumption is problematic because these two variables M_{x^*} and Y_{xM_x} can never be observed together.[28,51] Recent research proposed identification criteria[52] and effect decomposition[28] aimed at circumventing the violation of this fourth assumption. However, this issue is beyond the scope of this article. In addition, we assume no selection bias and measurement error. Under the above assumptions, different types of effect can be nonparametrically identified and estimated using the

empirical analogs listed in Table 2.2 (right column). For the current paper, we adopted a fully parametric approach, i.e., positing parametric (regression) models for each expectation in these empirical analogs to estimate their parameters from the observed data and then integrating over the covariate and/or mediator distribution.[53] We further assumed no model misspecification.

2.3.3 G-computation

G-computation algorithm was first introduced by Robins in 1986[39] to estimate the causal effect of a time-varying exposure in the presence of time-varying confounders that are affected by exposure, a scenario where traditional regression-based methods would fail. In recent years, several didactic examples were given in the literature,[41,54,55] promoting the use of this causal analytic technique. Increasingly, more studies have applied this methodology in estimating the effect of dynamic treatment regimes[56,57] or projecting the impact of hypothetical interventions.[58–62] In a simple setting with a single-time exposure and an outcome, g-computation can be seen as the generalization of standardization. Accessible examples of g-computation with detailed discussion of its strengths and limitations can be found elsewhere.[33,40,41,59]

2.3.4 G-computation steps in causal mediation analysis

The g-computation steps have been summarized in Figure 2.1. Step 1 involves obtaining the parameters of the assumed covariate distributions, and fitting the assumed models for the mediator (*M* model) and the outcome (*Y* model) using observed data. The key covariates needed for the *M* model are confounders of the exposure-mediator relation. The key covariates needed for the *Y* model are confounders of the exposure-outcome relation and mediator-outcome relation. To avoid simulating the covariate set $\boldsymbol{Z}$, one can replace steps 1a and 2a with resampling

(Figure 2.1). Step 2 entails simulating the covariate set $\boldsymbol{Z}$ (step 2a), an exposure intervention variable X (step 2b), and the potential mediator (step 2c), and outcome Y (step 2d) sequentially for J (J can be as large as computationally feasible) copies of the original sample. The simulation repetition done here is to reduce Monte Carlo simulation error. The intervention variable X should be distinguished from the observed exposure variable as the intervention X and the simulated covariates are marginally independent of each other. In step 2c, we simulate each potential mediator as a function of the simulated covariates and intervention X from the previous steps (2a and 2b), using the parameters obtained from the M model in step 1b. Similarly, in step 2d, we simulate a potential outcome variable that corresponds to each specific type of effect as a function of the simulated covariates, exposure intervention, and mediator from the previous steps, using the parameters obtained from the Y model in step 1c. Step 3 involves regressing each different potential outcome variable on the intervention variable X to obtain estimates of each marginal effect using the pooled data with J copies of the original sample. Repeat step 2-3 on K (usually 200 or more) bootstrapped samples taken at random with replacement from the original data. The Wald type 95% confidence interval (CI) was calculated as: point estimate $\pm$ 1.96 $\times$ SD, where SD was the standard deviation of the K resultant point estimates from the final regression in the third step.

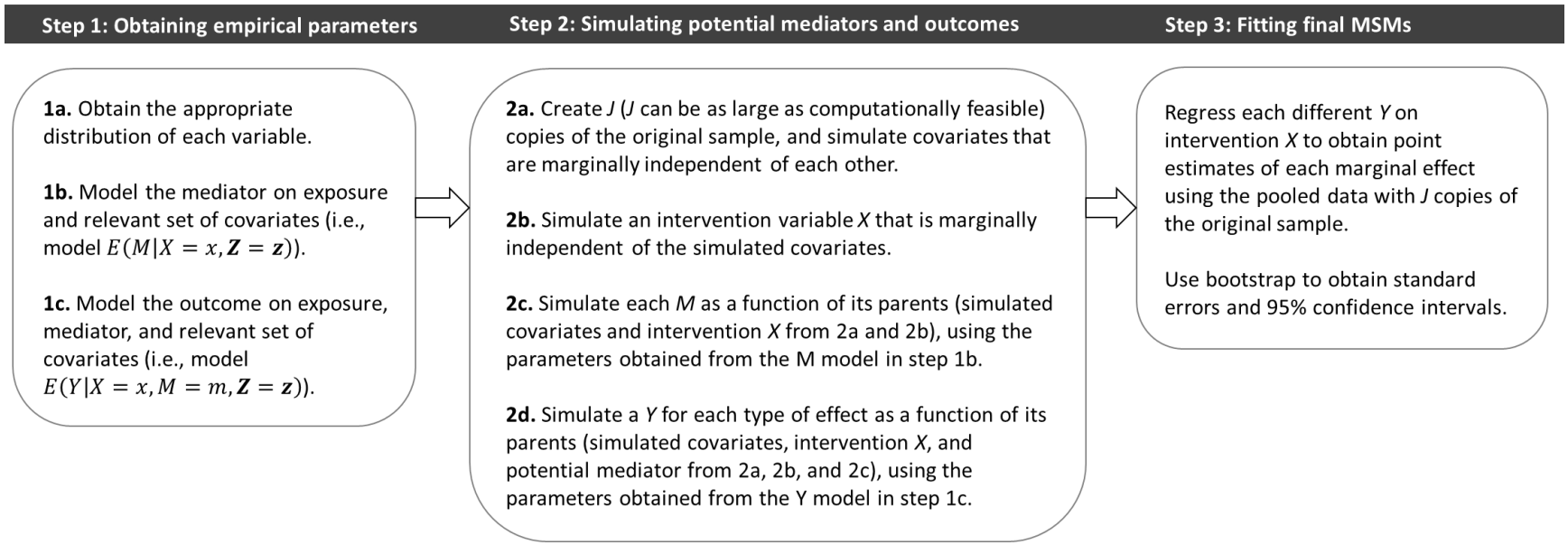

Figure 2.1 Steps for G-computation marginal structural model (G-comp MSM) in causal mediation analysis. Let Z denote a set of covariates, X denote exposure, M denote mediator, and Y denote outcome of interest. Variables in step 1 are observed variables whereas var

2.4 Illustration

We used a directed acyclic graph (DAG)[24] to represent the data generating process for the illustration example (Figure 2.2). We used the India sample from the World Health Survey (WHS).[63] We used all the observed covariates to simulate exposure smoking, mediator body mass index (BMI, 5-unit increase), and outcome composite health score (0-100) sequentially according to the data generating process shown in Figure 2.2. Covariates for confounding control included age, gender, education, urbanicity, and depression. Detailed description for generating this partially simulated data set can be found in the Appendix. In the illustrative example, we will focus on the interpretations for PDE and TIE from the most common decomposition used in epidemiology and the interpretations for component effects based on the recently introduced 4-way decomposition. Since smoking was binary, we used 1 to represent "yes" and 0 "no". A second illustrative example using binary exposure, mediator and outcome was included in the Appendix. All analyses were done using SAS version 9.4 (SAS Institute Inc., Cary, NC).

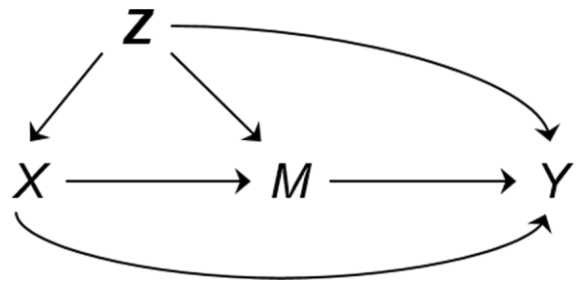

Figure 2.2 DAG representing the data generating process for the first illustrative example. *X*, *M*, and *Y* represent the exposure smoking, mediator body mass index, and outcome composite health score. **Z** represents a set of exposure-mediator, exposure-outcome and mediator-outcome confounders that includes age, gender, education, urbanicity, and depression.

To estimate different component effects of smoking on health, we implemented the following steps:

Step 1. Obtaining empirical parameters

(1a) We obtained the marginal expectation of each variable except the outcome, and the standard deviation for the continuous age variable.

(1b) The mediator BMI was regressed on smoking, age, gender, education, urbanicity and depression to obtain the regression coefficients and root mean square error (RMSE) for the linear *M* model.

(1c) The outcome, overall health score, was then regressed on smoking, BMI, smoking × BMI, age, gender, education and depression to obtain the regression coefficients and RMSE for the linear *Y* model.

Step 2. Simulating the potential mediators and outcomes

(2a) We created 1000 copies of the original sample and simulated age, gender, education, urbanicity, and depression that followed the same distribution as the observed variables.

(2b) We simulated a smoking intervention variable (*X*) that followed the observed smoking prevalence but was marginally independent of all simulated covariates.

(2c) We simulated each potential BMI variable as a function of the smoking intervention, age, gender, education, urbanicity and depression (Table 2.4, simulating *M*), using the regression coefficients and RMSE from the *M* model fit in (1b).

(2d) We simulated a potential health score variable for each type of effect as a function of the smoking intervention, potential BMI from (2c), product term between smoking intervention and

potential BMI, age, gender, education and depression (Table 2.4, simulating *Y*), using the regression coefficients and RMSE from the *Y* model fit in (1c).

We will use PDE and TIE as examples to explain steps (2c) and (2d) further. For binary exposure, recall that PDE compares X=1 to X=0 while setting the mediator M to M_0 (natural value under reference exposure). In (2c), we simulated the potential BMI variable (M_0) as a function of non-smoking (setting X=0 in the equation for simulating M) and other determinants of BMI. Next, we simulated the potential health score variable (Y_{PDE}) as a function of the smoking intervention variable (X), potential BMI variable (M_0) from (2c), and other determinants of health in (2d). On the other hand, TIE compares mediator level M_1 to M_0 while setting exposure to index level 1. In (2c), we simulated the potential BMI variable (M_x) as a function of smoking intervention (setting $X=x$ in the equation for simulating M) and other determinants of BMI. Then, we simulated the potential health score variable (Y_{TIE}) as a function of smoking (setting X=1 in the equation for simulating Y), potential BMI variable (M_x) from (2c), and other determinants of health in (2d). In this way, the potential BMI variable (M_x) transmitted the effect of smoking intervention to health.

Step 3. Fitting final marginal structural models (MSMs)

We regressed each different potential health score variable on the smoking intervention to obtain point estimates of each marginal effect using the pooled sample. We repeated step 2-3 on 200 bootstrapped samples of the same size taken at random with replacement from the original data to obtain Wald type 95% CIs.

Table 2.4 Equations used to simulate potential mediators and outcomes in step 2 of the G-computation marginal structural model for the illustrative example.[a]

M model and Y model from step 1[b]:

$E(M|x, \mathbf{z}; \alpha) = \alpha_M + \alpha_X \cdot x + \alpha_{\mathbf{Z}} \cdot \mathbf{z}$

$E(Y|x, m, \mathbf{z}; \beta) = \beta_Y + \beta_X \cdot x + \beta_M \cdot m + \beta_{XM} \cdot x \cdot m + \beta_Z \cdot \mathbf{z}$

Effect	**Simulating M^c**		**Simulating Y^c**	
TE[d]	M_x	$= \alpha_M + \alpha_X \cdot x + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$[e]	Y_{TE}	$= \beta_Y + \beta_X \cdot x + \beta_M \cdot m_x + \beta_{XM} \cdot x \cdot m_x + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$[e]
PDE	M_0	$= \alpha_M + \alpha_X \cdot 0 + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$	Y_{PDE}	$= \beta_Y + \beta_X \cdot x + \beta_M \cdot m_0 + \beta_{XM} \cdot x \cdot m_0 + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
TIE	M_x	$= \alpha_M + \alpha_X \cdot x + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$	Y_{TIE}	$= \beta_Y + \beta_X \cdot 1 + \beta_M \cdot m_x + \beta_{XM} \cdot 1 \cdot m_x + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
TDE	M_1	$= \alpha_M + \alpha_X \cdot 1 + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$	Y_{TDE}	$= \beta_Y + \beta_X \cdot x + \beta_M \cdot m_1 + \beta_{XM} \cdot x \cdot m_1 + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
PIE	M_x	$= \alpha_M + \alpha_X \cdot x + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$	Y_{PIE}	$= \beta_Y + \beta_X \cdot 0 + \beta_M \cdot m_x + \beta_{XM} \cdot 0 \cdot m_x + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
$CDE_{M=m^*}$	m^*	$= 4.8$[f]	Y_{CDE}	$= \beta_Y + \beta_X \cdot x + \beta_M \cdot m^* + \beta_{XM} \cdot x \cdot m^* + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
CDE_{sto}	M'	$= E(M) + \varepsilon_M$	Y_{CDEsto}	$= \beta_Y + \beta_X \cdot x + \beta_M \cdot m' + \beta_{XM} \cdot x \cdot m' + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
RIE	M_0	$= \alpha_M + \alpha_X \cdot 0 + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$	Y_{RIE}[g]	$= \beta_Y + 0 \cdot x + 0 \cdot m_x + \beta_{XM} \cdot x \cdot (m_0 - m^*) + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
MIE	M_{med}[h]	$= \alpha_X \cdot x + \varepsilon_M$	Y_{MIE}[g]	$= \beta_Y + 0 \cdot x + 0 \cdot m_x + \beta_{XM} \cdot x \cdot m_{med} + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$
PAI	M_1	$= \alpha_M + \alpha_X \cdot 1 + \alpha_{\mathbf{Z}} \cdot \mathbf{z} + \varepsilon_M$	Y_{PAI}[g]	$= \beta_Y + 0 \cdot x + 0 \cdot m_x + \beta_{XM} \cdot x \cdot (m_1 - m^*) + \beta_Z \cdot \mathbf{z} + \varepsilon_Y$

[a] Exposure: smoking (1=yes, 0=no); mediator: body mass index (5-unit increase); outcome: composite health score; covariates: age, gender, education, urbanicity, depression.

[b] Variables used to fit the M model and Y model in step 1 were observed variables.

[c] Lower case "x", "$\mathbf{z}$", and "m_x" represented specific values of the random variables intervention "*X*", simulated covariate set "**Z**", and potential mediator "M_x" respectively and the values can differ for different individuals. Intervention "*X*" was independent of the simulated covariate set "**Z**".

[d] TE: total effect, PDE: pure direct effect, TIE: total indirect effect, TDE: total direct effect, PIE: pure indirect effect, CDE: controlled direct effect (standard), CDE_{sto}: stochastic controlled direct effect, RIE: reference interaction effect (referred to as "INT_{ref}" by VanderWeele), MIE: mediated interaction effect (referred to as "INT_{med}" by VanderWeele), PAI: portion attributable to interaction.

[e] The root mean square error (RMSE) from the M model and the Y model in step 1 respectively.

[f] The mediator was fixed at 4.8 (BMI=24).

[g] To simulate Y_{RIE}, Y_{MIE}, and Y_{PAI}, we assigned zero for the coefficients for random variables intervention "*X*" and the potential mediator "M_x" but not the coefficients for the product term between these two variables to mimic "de-activating" the direct and indirect path from *X* to *Y*, leaving only a specific type of "interaction" between *X* and *M* to transmit the effect of *X* to *Y*.

[h] The mediated interaction captures the interaction between *X* and a version of *M* that is due to *X* only. Thus, to simulate M_{med}, the mediator *M* responds to no other determinants of *M* but *X*.

Results from the illustrative example were presented in Table 2.5. This example is for illustration purpose and thus the results were not intended for quotation. Estimates followed the effect decompositions as described in Table 2.1. Smoking had an overall negative impact on health (TE: –0.96, 95% CI: –1.79, –0.13), but the majority of this impact was through pathways other than changing BMI (PDE: –0.70, 95% CI: –1.54, 0.14). When BMI was fixed at 24 for everyone, smoking did not appear to affect health directly (CDE: 0.27, 95% CI: –0.95, 1.50). In a hypothetical intervention where BMI was no longer affected by smoking and other covariates, smoking still had a negative impact on health (CDE_{sto}: –0.81, 95% CI: –1.63, 0.02). The impact of smoking on health that was due to interaction with BMI (RIE: –0.99, 95% CI: –1.71, –0.26) was much larger than the part that was due to both mediation and interaction (MIE: –0.14, 95% CI: –0.27, –0.01). The presence of such smoking-BMI interaction contributed to the difference seen when comparing CDE to PDE and CDE_{sto}.

Table 2.5 Effect estimate (95% Confidence Interval) for the illustrative example[a] using g-computation of marginal structure models[b] (N=5326).

Effect	***b* (95% CI)**
Total Effect (TE)	-0.96 (-1.79, -0.13)
Pure Direct Effect (PDE)	-0.70 (-1.54, 0.14)
Total Indirect Effect (TIE)	-0.26 (-0.42, -0.10)
Total Direct Effect (TDE)	-0.87 (-1.69, -0.04)
Pure Indirect Effect (PIE)	-0.12 (-0.22, -0.02)
Controlled Direct Effect[c] (CDE)	0.27 (-0.95, 1.50)
Stochastic Controlled Direct Effect[d] (CDE_{sto})	-0.81 (-1.63, 0.02)
Reference Interaction Effect (RIE)	-0.99 (-1.71, -0.26)
Mediated Interaction Effect (MIE)	-0.14 (-0.27, -0.01)
Portion Attributable to Interaction (PAI)	-1.14 (-1.96, -0.32)

[a] Exposure: smoking (1=yes, 0=no); mediator: body mass index (BMI, 5-unit increase); outcome: composite health score; covariates: age, gender, education, urbanicity, depression.

[b] Effect estimates were based on 1000 simulation replicates and confidence intervals were based on 200 bootstrapped samples paired with 200 simulation replicates.

[c] The mediator was fixed at 4.8 (BMI=24) for every individual in the sample.

[d] The mediator was allowed to obtain a certain distribution of M' where the mediator has the mean (mean BMI=20.7) and the variability of the observed mediator but was independent of its determinants (exposure and covariates).

2.5 Discussion

In this article we demonstrated the utility of parametric g-computation in estimating various marginal effects under different and detailed effect decompositions as well as stochastic controlled direct effects, using Monte Carlo simulations in standard statistical software. To our knowledge, this is the first use of g-computation for 3- and 4- way decomposition of effects recently introduced by VanderWeele.[43] Our approach yielded similar results as those obtained from VanderWeele's approach[43] (appendix Table A 2.3). However, marginal (standardized) effect measures obtained via g-computation approach are not conceptually equal to the conditional (on covariates) effect measures obtained from the latter approach and results from the two approaches may differ.[64] Alternative imputation[65] or simulation[35] based methods are also available for common mediation parameters.

G-computation has several strengths. It uses models for the outcome and the mediator, which produces more efficient estimates (with narrower confidence intervals) than the weighting approaches that use models for the mediator and/or the exposure.[65,66] It can be used to estimate various types of effect of interest, incorporate nonlinearities and exposure-covariate and mediator-covariate interactions, and deal with general types of outcome, exposure and mediator. This simulation-based approach can be used to estimate various effects on both difference and ratio scales.

However, g-computation is not without its limitations. The parametric g-computation method applied in mediation settings, like the general g-formula, relies on a correctly

specified model for the outcome. For natural effects specifically, it additionally requires that the model for the mediator is correctly specified, as with other approaches published previously.[66–70] When such parametric distribution for *M* is in doubt, a distribution-free approach with regard to the mediator[71] can be used. Alternative approaches are to use non-parametric g-computation method that combines bootstrapping and simulation as suggested in Imai et al.[35] or implement a doubly robust estimator that requires at least one of the model for exposure and mediator being correctly specified.[65] In addition, the computation time depends on the sample size and the number of covariates. As these two numbers increase, a random subset of the sample can be selected to perform the Monte Carlo simulation.[33]

G-computation in mediation analysis, especially parametric g-formula implemented via Monte Carlo simulation, can be seen as a special application of the longitudinal time-varying g-computation formula.[32,39] In this case, both post-baseline exposure and confounders that are affected by exposure can be seen as mediators. The steps are similar in simulating the baseline confounders and exposure intervention first and then the consequences of the exposure following the data structure represented in a specific DAG, though additional assumptions are needed in the mediation setting. G-computation as a unifying and flexible framework will gain popularity with the increase in applications of mediation analysis to answer mechanistic questions about either contextual or individual level causes.[36–38]

By showing the steps for g-computation in estimating different quantities of interest in causal mediation analysis, we hope to encourage a wider audience of applied researchers to implement this framework, using software packages of their choice. Given the growing interest in adopting and applying complex systems approaches to examine complex disease etiologies, this method, especially with its simulation component, will be an important intermediate step towards this journey.[72]

2.6 Appendix

Description of the partially simulated data set

The simulated data set is a combination of variables from the India sample in the World Health Survey and simulated variables using the Monte Carlo method. The World Health Survey (WHS) is a cross-sectional survey administered by the World Health Organization (WHO) in 70 countries between 2002 and 2004.[63] For the current illustration, we used data from the India sample and restricted to participants who were at least 25 years old and with complete data on the exposure, mediator, outcome, and relevant covariates. We also excluded participants with extreme height and weight using the same method published previously.[73] The resulting sample size is 5326 for the first example and is 6527 for the second example. We used all the observed covariates as they were from the original data to simulate exposure *X*, mediator *M*, and outcome *Y* sequentially according to the data generating process shown in Figure 2.2 in the main manuscript, using the *rand* function in SAS. All variables used for each example and the data generating protocol are listed in Table A 2.1. The magnitude between variables was taken from the real data (the empirical estimates based on regression). The equations used to generate *X*, *M*, and *Y* with the realized coefficients were presented in the lower part of Table A 2.1. Despite the cross-sectional nature of the observed dataset, we imposed causal direction between variables *X*, *M*, and *Y* in our partially simulated dataset via simulation.

Variables and simulation protocol for illustrative examples

Table A 2.1 Variables and simulation protocol for exposure *X*, mediator *M*, and outcome *Y* in two illustrative examples.

Variable	Example 1: binary *X*, continuous *M* and *Y* (N=5326)	Example 2: binary *X*, *M* and *Y* (N=6527)
X	Ever smoking (yes versus no), *P*(smoking) = 0.37	Urbanicity (urban versus rural), *P*(urban) = 0.30
M	Body mass index (BMI, 5-unit increase), mean = 4.15 (SD = 0.67)	Overweight (yes: BMI>=25 versus no BMI<25), *P*(overweight) = 0.10
Y	Composite health score (0-100), mean = 76.03 (SD = 15.06)	Ever diagnosed with diabetes (yes versus no), *P*(diabetes) = 0.037
X-M confounders	Urbanicity (urban versus rural) , *P*(urban) = 0.29	None
X-Y confounders	None	None
M-Y confounders	Depression (yes versus no), *P*(depression) = 0.11	None
X-M, X-Y, and M-Y confounders	Age (continuous, re-centered to 25 years,10-year increase), mean = 1.83 (SD = 1.38) Gender (female versus male), *P*(female) = 0.50 Education (primary school completed and beyond versus less than primary school), *P*(primplus) = 0.48	Age (continuous, re-centered to 25 years,10-year increase), mean = 1.79 (SD = 1.37) Gender (female versus male), *P*(female) = 0.51 Education (primary school completed and beyond versus less than primary school), *P*(primplus) = 0.48
Equation for *X*	$\mathrm{X} \sim \mathrm{B}\,(1,\ 1/\,(1+\exp(-(0.71+0.084\cdot \mathrm{age}-2.17\cdot \mathrm{gender}-0.25\cdot \mathrm{urbanicity}-0.80\cdot \mathrm{education}))))$	$\mathrm{X} \sim \mathrm{B}\,(1,\ 0.10-0.015\cdot \mathrm{age}+0.082\cdot \mathrm{gender}+0.27\cdot \mathrm{education})$
Equation for *M*	$\mathrm{M} \sim \mathrm{N}\,(3.93-0.090\cdot x+0.051\cdot \mathrm{age}-0.11\cdot \mathrm{gender}+0.21\cdot \mathrm{urbanicity}+0.12\cdot \mathrm{education}-0.078\cdot \mathrm{depression}, 0.65)$	$\mathrm{M} \sim \mathrm{B}\,(1,\ 0.016+0.076\cdot x+0.013\cdot \mathrm{age}+0.037\cdot \mathrm{gender}+0.047\cdot \mathrm{education})$
Equation for *Y*	$\mathrm{Y} \sim \mathrm{N}\,(79.11-7.52\cdot x+1.23\cdot m+1.62\cdot xm-3.80\cdot \mathrm{age}-2.90\cdot \mathrm{gender}+3.11\cdot \mathrm{education}-8.01\cdot \mathrm{depression}, 13.51)$	$\mathrm{Y} \sim \mathrm{B}\,(1,\ 0.016+0.028\cdot x+0.022\cdot m+0.058\cdot xm+0.010\cdot \mathrm{age}-0.013\cdot \mathrm{gender}+0.011\cdot \mathrm{education})$

Detailed description for g-computation steps for the second illustrative example

To estimate different component effects of smoking on health, we implemented the following steps:

Step 1. Obtaining empirical parameters

(1a) We obtained the marginal expectation of each variable except the outcome, and the standard deviation for the continuous age variable.

(1b) The mediator overweight indicator was regressed on urbanicity, age, gender and education to obtain the regression coefficients for the linear risk *M* model.

(1c) The outcome, diabetes indicator, was then regressed on urbanicity, overweight indicator, urbanicity× overweight indicator, age, gender and education to obtain the regression coefficients for the linear risk *Y* model.

Step 2. Simulating the potential mediators and outcomes

(2a) We created 1000 copies of the original sample and simulated age, gender and education that followed the same distribution as the observed variables.

(2b) We simulated an urbanicity intervention variable (*X*) that followed the observed prevalence of living in urban area but was marginally independent of all simulated covariates.

(2c) We simulated each potential overweight indicator as a function of the urbanicity intervention, age, gender and education, using the regression coefficients from the M model fit in (1b).

(2d) We simulated a potential diabetes indicator for each type of effect as a function of the urbanicity intervention, potential overweight indicator from (2c), product term

between urbanicity intervention and potential overweight indicator, age, gender and education, using the regression coefficients from the Y model fit in (1c).

Step 3. Fitting final marginal structural models (MSMs)

We regressed each different potential diabetes indicator on the urbanicity intervention to obtain point estimates of each marginal effect using the pooled sample. We repeated step 2-3 on 200 bootstrapped samples of the same size taken at random with replacement from the original data to obtain Wald type 95% CIs.

Results from the second illustrative example

Table A 2.2 Effect estimate (95% Confidence Interval) for the second illustrative example[a] using g-computation of marginal structure models[b] (N=6527).

Effect	RD (95% CI)
Total Effect (TE)	-0.039 (-1.788, -0.130)
Pure Direct Effect (PDE)	-0.033 (-1.536, 0.135)
Total Indirect Effect (TIE)	-0.006 (-0.415, -0.098)
Total Direct Effect (TDE)	-0.037 (-1.690, -0.041)
Pure Indirect Effect (PIE)	-0.002 (-0.220, -0.023)
Controlled Direct Effect[c] (CDE)	-0.028 (-0.953, 1.502)
Stochastic Controlled Direct Effect[d] (CDE_{sto})	-0.034 (-1.633, 0.021)
Reference Interaction Effect (RIE)	-0.005 (-1.708, -0.262)
Mediated Interaction Effect (MIE)	-0.005 (-0.270, -0.009)
Portion Attributable to Interaction (PAI)	-0.009 (-1.960, -0.317)

[a] Exposure: urbanicity (1=urban, 0=rural); mediator: overweight (1=yes, 0=no); outcome: diabetes (1=yes, 0=no); covariates: age, gender, education.
[b] Effect estimates were based on 1000 simulation replicates and confidence intervals were based on 200 bootstrapped samples paired with 200 simulation replicates.
[c] The mediator is fixed at 0 (non-overweight) for every individual in the sample.
[d] Allow the mediator to obtain a certain distribution of M' where the mediator has prevalence of the observed mediator but is independent of its determinants (exposure and covariates).

Results from the first illustrative example using the approach by VanderWeele (2014)[43]

Table A 2.3 Effect estimate (95% Confidence Interval) for the first illustrative example[a] using VanderWeele's approach[b] (N=5326).

Effect	*b* (95% CI)
Total Effect (TE)	-0.99 (-1.84, -0.14)
Pure Direct Effect (PDE)	-0.73 (-1.58, 0.12)
Total Indirect Effect (TIE)	-0.26 (-0.40, -0.12)
Total Direct Effect (TDE)	-0.88 (-1.73, -0.03)
Pure Indirect Effect (PIE)	-0.11 (-0.19, -0.03)
Controlled Direct Effect[c] (CDE)	0.27 (-0.89, 1.44)
Stochastic Controlled Direct Effect[d] (CDE_{sto})	--
Reference Interaction Effect (RIE)	-1.01 (-1.72, -0.30)
Mediated Interaction Effect (MIE)	-0.15 (-0.27, -0.02)
Portion Attributable to Interaction (PAI)	-1.15 (-1.97, -0.34)

[a] Exposure: smoking (1=yes, 0=no); mediator: body mass index (BMI, 5-unit increase); outcome: composite health score; covariates: age, gender, education, urbanicity, depression.
[b] All effect estimates were obtained at the mean value of the covariates.

Chapter 3. The Impact of Human Development on Individual Health: a Causal Mediation Analysis Examining Pathways Through Education and Body Mass Index

3.1 Abstract

Introduction: Improved health is known to predict national economic growth and development. Yet, the reciprocal effect of development on individual health is rarely examined. This study examined the impact of human development on individual health and the possible mediating roles of education and body mass index (BMI).

Methods: We analyzed data on 109,448 participants aged 25 or older from 42 low- and middle-income countries that participated in the World Health Survey 2002-2004. These data were augmented with the 1990 human development index (HDI). The outcome was a health score based on measures from eight health state domains, years of schooling was used as education indicator, and BMI was calculated from self-reported height and weight. We used modern causal mediation analytical techniques implemented as linear mixed models with random intercepts to analyze the multilevel data.

Results: Below a reference HDI level of 0.48, HDI was negatively associated with good health (total effect at HDI of 0.23: b = –3.44, 95% CI: –6.39, –0.49 for males and b = –5.16, 95% CI: –9.24, –1.08 for females) but was positively associated with good health above this reference level (total effect at HDI of 0.75: b = 4.16, 95% CI: –0.33, 8.66 for males and b = 6.62, 95% CI: 0.85, 12.38 for females). A small positive effect of HDI on health was found via education across reference HDI levels (b ranging from 0.24 to 0.29 for males and 0.40 to 0.49 for females) but not pathways involving BMI only.

Conclusion: Human development has a non-linear effect on individual health, and mainly through pathways other than individual level education and BMI. Modern causal mediation analysis seems promising for examining and decomposing contextual health effects of human development in global comparative studies.

3.2 Introduction

Better health is central to human happiness and well-being. It also makes an important contribution to economic progress. Accumulating evidence showed that increased investment in health translated into improved individual productivity, longer work life, and in turn, country's income.[74] Besides the old adage of "health before wealth", it is important to learn the other side of the coin: the reciprocal loop from development to health, and more importantly, the underlying mechanisms. Huge amount of evidence demonstrated how individual educational attainment affects our health and efforts have been made to reduce the disparity in health due to one's socioeconomic achievement. Evidence is relatively scant on questions such as "how the social environment we live in projects what we can achieve and how we behave, and then in turn shapes our health?" A person may have much higher chance of staying at the high tier of social ladder and adopt a healthy lifestyle when he or she lives in a country where higher education is covered by government than a country where basic schooling is not ensured. The health benefits of education were found to be dependent on a country's level of human development[75] or country context[76]. Indeed, we cannot discuss a person's health without considering the 'contextual web' the person lives in, as suggested in previous literature.[77]

Adoption of health behaviors is one of the important potential mechanisms connecting country development, education and health. Obesity, a crucial indicator closely related to individual health behavior, has been linked to various adverse health outcomes.[73,78,79] Studies have shown that obesity prevalence is much higher among people with low socioeconomic status (SES) in developed countries but the burden of obesity shifts towards the poor rapidly in the developing world as their economies grow.[80] As countries develop and globalize, foreign fast-food companies replace the farmer's market in local areas, and skyscrapers replace walkable areas, exposing citizens to obesogenic environments. Interesting questions are to what extent the macro environment we live in shapes our health via influencing our education, and to what extent it shapes our health via affecting our weight status. This study aimed at addressing these two issues. More specifically, we used standardized global data to investigate (1) the impact of country's human development level as measured by human development index on individual health, and (2) the mediating role of both education and body mass index in the relation between human development and health.

3.3 Method

Study sample and variables

We used data from 49 low- and middle-income countries (LMICs) collected by the cross-sectional World Health Survey (WHS). Conducted by the WHO from 2002 to 2004, the WHS used standardized methodology that provided a basis for examining individual health measures across countries.[81] The details for study design and methods of the WHS can be found in the appendices and elsewhere.[81,82]

Outcome

Individual were asked to report their perceived difficulties on eight health state domain (two questions per domain) using 5-point Likert scale questions: mobility, self-care, pain and discomfort, cognition, interpersonal activities, vision, sleep and energy, and affect.[81] The health state measures have been extensively tested[83] and have been shown to have good consistency and reliability.[82] We performed factor analysis using polychoric correlations to account for the covariance structure and ordinal nature of the responses to individual questions. Similar to a previous study,[82] we chose one factor solution based on the high eigenvalue of the first factor (8.85, 73% as a cumulative percentage of the variance explained) and the high communalities of the original variables (between 0.36 and 0.69). Then, we used the principal component method for factor extraction and the regression scoring method to obtain the factor scores. The factor score was rescaled with 0 indicating worst health and 100 indicating best health.

Exposure

Human Development Index (HDI) is chosen as a measure of the national socioeconomic environment for human development in a country. HDI is a unit-free index between 0 and 1 that is calculated for each country based on life expectancy at birth, adult literacy rate, combined gross enrollment ratio for primary, secondary and tertiary education, and Gross Domestic Product per capita (GDP/c). In this study, we used HDI reported for 1990[84] and rescaled the score to range from 0 to 10. We lag the HDI for more than 10 years to capture its effect on shaping individual's education and assumed that HDI from 1990 is a good indicator for the country's socioeconomic environment for periods before 1990.

Mediators

Individual-level education was measured by the years of schooling (including higher education).

Body mass index (BMI) was defined as an individual's self-reported weight (kg) divided by self-reported height squared (m^2). We excluded participants with height less than 1.22 m (n=1174) or greater than 2.11 m (n=27), and participants with weight that was 3 SDs above (n=821) or 2 SDs below (n=382) the crude sample mean of 63.6 kg. We further excluded individuals with BMI less than 14 kg/m^2 (n=252).

Confounders

Potential confounders are WHO region, individual age and sex. In sensitivity analyses, we further considered potential confounders of the BMI-health relationship that were possibly influenced by education and/or HDI: living in urban areas, unemployment, marital status, and health behaviors such as smoking, alcohol use, and physical activity.

Conceptual framework

We used a directed acyclic graph (DAG)[24] to represent our assumptions about the data generating process (Figure 3.1). Let Y denote individual health score, X the country's human development index, M the mediator of interest: education in scenario 1 and body mass index in scenario 2, L the exposure-induced mediator-outcome confounder: education in scenario 2. Let $Y(x, M(x))$ denote the potential Y had X been set to x, M been allowed to attain its natural value under intervention $X = x$. Let $Y(x, L(x), M(x, L(x)))$ denote the potential Y had X been set to x, L been set to the natural value under $X = x$, M been set to the natural value under $X = x$ and $L = L(x)$. Let x_1 (index) and x_0 (reference)

represent two exposure values we wish to compare. Table 3.1 (left column) shows the effect definition when education is the mediator of interest (scenario 1) or when BMI is the mediator of interest (scenario 2). In scenario 1, the pure direct effect captured the impact of human development level on individual health through pathways other than individual's education (Figure 3.2a) whereas the total indirect effect measured such impact through education (Figure 3.2b). In examining the mediating role of BMI, a consequence of individual education, we further decomposed the pure direct effect of HDI on health into: 1) the HDI effect through BMI but not education (i.e. the BMI-path-specific effect as presented in Figure 2c) and 2) the natural direct effect of HDI on health through neither education nor BMI. Detailed assumptions necessary for effect identification are listed in Appendix.

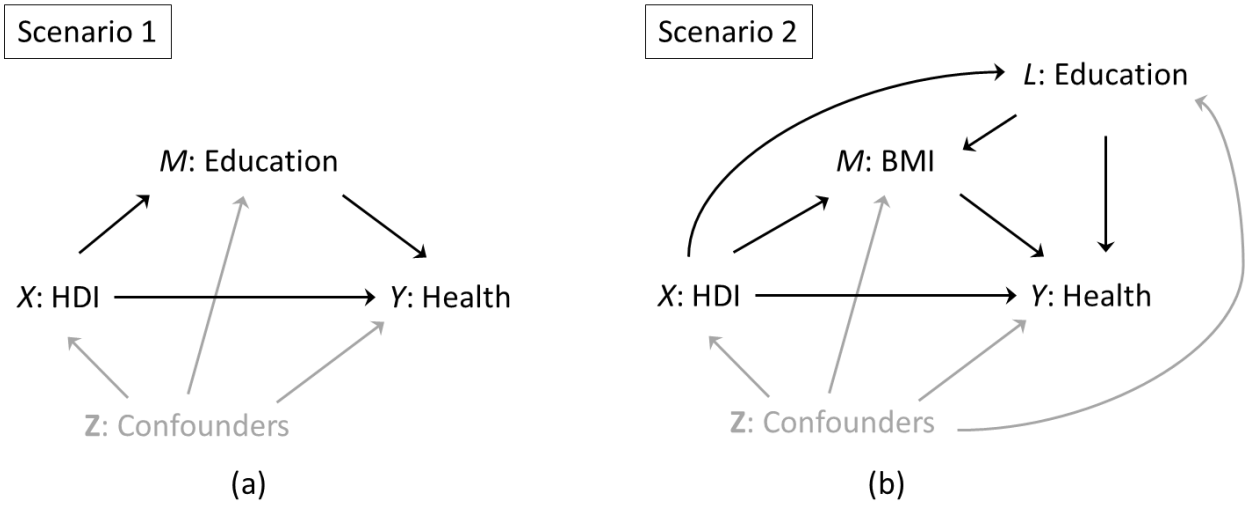

Figure 3.1 DAG depicting hypothesized data generating process involving a single exposure, mediator, and outcome. (a) *X*: human development index (HDI) as contextual exposure, *M*: education, *Y*: health, Z: a set of measured confounders of the *X*-*M*, *X*-*Y*, and *M*-*Y* relationships at group or individual level. (b) *X*: HDI, *M*: education, *Y*: health, *L*: confounder of the *M*-*Y* relationship that is a consequence of *X*, called "intermediate confounder" or "endogenous confounder".

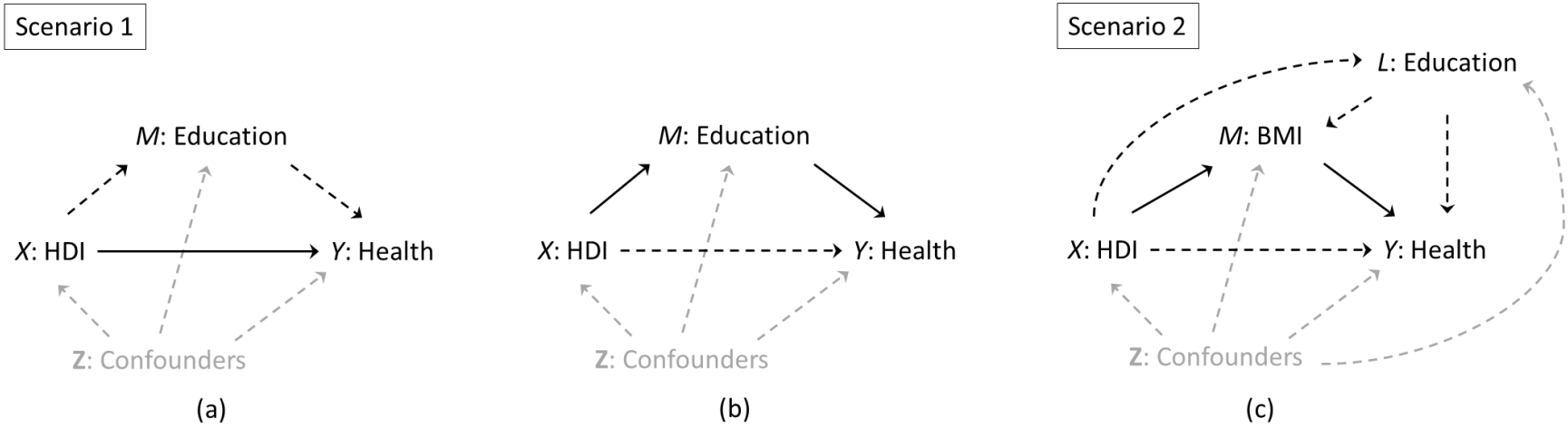

Figure 3.2 Graphical representation (solid lines) of pure direct effect of HDI on health (a), total indirect effect of HDI via education (b), and natural indirect effect of HDI via BMI only (c). In scenario 1, the BMI path-specific effect was incorporated in the pure direct effect.

Statistical analysis

We used appropriate descriptive statistics to summarize the characteristics of the participants by sex and WHO region. Under the assumptions of general consistency, conditional exchangeability (no uncontrolled confounding), and positivity[28], we can estimate the effects defined in Table 3.1 for continuous exposure, mediator, and outcome using the empirical expressions listed in Table 3.1 (right column). To account for the clustering within country, we used generalized linear mixed models with country-specific random intercept for each model of health score, BMI, and education. We estimated point mean differences and their corresponding 95% confidence intervals (95% CIs) for males and females separately. All analyses were conducted in SAS 9.4 (SAS Institute Inc., Cary, North Carolina, USA).

Sensitivity analysis

For scenario 1, we relaxed the sample restriction criteria to individuals with complete information on HDI, education, and health score (N=148,679) and re-estimated the pure direct effect, total indirect effect and total effect. In scenario 2, we further examined the robustness of our estimates in the presence of intermediate confounders **V** (affected by HDI, education or both) of the BMI-health relationship. We used g-estimation technique to create a confounding-free outcome variable where this new outcome variable is independent of **V** conditional on HDI, education, BMI, and other covariates (the set **Z**). Details can be found in the appendix.

Table 3.1 Effect definition and empirical expressions applied to the World Health Survey 2002-2004.

Effect	**Counterfactual definition**[a]	**Empirical expression**[b]
Scenario 1	*X*: HDI, *M*: education years, *Y*: health score $E(M\|x, z^M; \delta) = \delta_M + \delta_X \cdot x + \delta_Z \cdot z^M$ $E(Y\|x, m, z; \theta) = \theta_Y + \theta_X \cdot x + \theta_{X^2} \cdot x^2 + \theta_M \cdot m + \theta_{XM} \cdot x \cdot m + \theta_Z \cdot z$	
TE[c]	E{$Y(x_1, M(x_1))$ – $Y(x_0, M(x_0))$}	$\{\theta_X + \theta_M \cdot \delta_X + \theta_{XM} \cdot [\delta_M + \delta_X \cdot (x_1 - x_0) + \delta_Z \cdot z^M]\} \cdot (x_1 - x_0) + \theta_{X^2} \cdot ({x_1}^2 - {x_0}^2)$
PDE	E{$Y(x_1, M(x_0))$ – $Y(x_0, M(x_0))$}	$\{\theta_X + \theta_{XM} \cdot (\delta_M + \delta_X \cdot x_0 + \delta_Z \cdot z^M)\} \cdot (x_1 - x_0) + \theta_{X^2} \cdot ({x_1}^2 - {x_0}^2)$
TIE	E{$Y(x_1, M(x_1))$ – $Y(x_1, M(x_0))$}	$\{\theta_M \cdot \delta_X + \theta_{XM} \cdot \delta_X \cdot x_1\} \cdot (x_1 - x_0)$
Scenario 2	*X*: HDI, *L*: education years, *M*: BMI, *Y*: health score $E(L\|x, z^L; \alpha) = \alpha_L + \alpha_X \cdot x + \alpha_Z \cdot z^L$ $E(M\|x, z^M; \beta) = \beta_M + \beta_X \cdot x + \beta_L \cdot l + \beta_Z \cdot z^M$ $E(Y\|x, m, z; \gamma) = \gamma_Y + \gamma_X \cdot x + \gamma_{X^2} \cdot x^2 + \gamma_L \cdot l + \gamma_M \cdot m + \gamma_{XL} \cdot x \cdot l + \gamma_{XM} \cdot x \cdot m + \gamma_{LM} \cdot l \cdot m + \gamma_Z \cdot z$	
$NIE_{X\to M\to Y}$	E{$Y(x_1, L(x_0), M(x_1, L(x_0)))$ – $Y(x_1, L(x_0), M(x_0, L(x_0)))$}	$\beta_X \cdot \{\gamma_M + \gamma_{LM} \cdot \alpha_L + \gamma_{XM} \cdot x_1 + \gamma_{LM} \cdot \alpha_X \cdot x_0 + \gamma_{LM} \cdot \alpha_Z \cdot z^L\} \cdot (x_1 - x_0)$

[a] We used x_1 (index) and x_0 (reference) to denote the two exposure value that we wish to compare.

[b] The average effect were conditional on covariates **Z**=**z**. We used z^M, z^L and z to denote the set of covariates included in the model for *M*, *L*, and *Y* respectively.

[c] TE: total effect of HDI on health, PDE: pure direct effect, TIE: total indirect effect of HDI on health via education, $NIE_{X\to M\to Y}$: natural indirect effect of HDI on health via BMI but not education.

3.4 Result

Among 195,808 participants aged 25 years or older, 109,448 (55.9%) participants from 42 countries had complete information on all covariates. Country-specific sample size and characteristics are presented in Table A 3.1 of the appendix. Participants from Burkina Faso, Chad, Comoros, Ethiopia, Herzegovina, and Georgia (N=15,770) were excluded because of missing HDI.

Table 3.2 shows participant characteristics by sex and WHO region. HDI varied by WHO region, with the European region having the highest mean HDI (0.71) and the South-East Asia region the lowest (0.44). Participants from Europe were the oldest (mean age: 48.0 for males and 49.0 for females), most educated (mean years of schooling: 12.4 for males and 12.2 for females), had highest mean BMI values (25.7 for males and 25.9 for females), but reported the lowest health scores (86.4 for males and 82.2 for females). Participants from WHO regions other than Europe and the Americas had similar sex-specific mean age, education years, BMI, and health scores. Overall, females were less educated and reported poorer health. Descriptive table for participants with complete information on HDI, education, and health score only is included in the appendix (Table A 3.2), and revealed similar patterns.

Sex-specific mean differences in health score associated with a 0.1-unit increase in HDI at the median HDI level (Table 3.3) or at multiple reference HDI levels (Figure 3.3). At the median HDI level of 0.572 in scenario 1, increase in HDI was positively associated with better health in both males (b = 1.58, 95% CI: –0.61, 3.77) and females (b = 2.61, 95% CI: –0.09, 5.32). The majority of such impact was through pathways other than individual's education (male: b = 1.32,

95% CI: –0.87, 3.51; female: *b* = 2.18, 95% CI: –0.52, 4.88). A small positive indirect effect of HDI via education was seen in both males (*b* = 0.26, 95% CI: 0.17, 0.35) and females (*b* = 0.44, 95% CI: 0.28, 0.59). The BMI-path-specific effect of HDI was near null in both sexes (male: *b* = 0.016, 95% CI: –0.005, 0.037; female: *b* = –0.033, 95% CI: –0.077, 0.011). All types of effects of HDI on health depended on the reference value of HDI. Increase in HDI below a reference HDI level of 0.483 was negatively associated with good health (total effect at HDI of 0.232: *b* = –3.44, 95% CI: –6.39, –0.49 for males and *b* = –5.16, 95% CI: –9.24, –1.08 for females) but was positively associated with good health above this reference level (total effect at HDI of 0.747: *b* = 4.16, 95% CI: –0.33, 8.66 for males and *b* = 6.62, 95% CI: 0.85, 12.38 for females). This pattern for the total effect was also seen in pure direct effect. As HDI increases, total indirect effect of HDI via education decreased slightly, with effect estimates ranging from 0.32 to 0.35. Natural indirect effect via BMI only also decreased and became negative for reference HDI level above 0.483, with effect estimates ranging from –0.07 to 0.08. The effect size among females was larger than that among males.

Sensitivity analyses using less restricted sample revealed similar point estimates and narrower confidence intervals (Table A 3.3 and Figure A 3.1). Summary statistics for potential intermediate confounders are presented in the appendix (Table A 3.4). After accounting for these intermediate confounders (Figure A 3.2), effect estimates for the natural indirect effect via BMI only were similar as those from the main analyses (Figure A 3.3).

Table 3.2 Participant characteristics by WHO region, World Health Survey 2002-2004 (N=109,448).

Characteristics, mean (SD)	**Africa**	**The Americas**	**Eastern Mediterranean**	**Europe**	**South-East Asia**	**Western Pacific**	**All**
Male							
Total, N (%)	9873 (19.6)	16072 (31.8)	2710 (5.4)	4562 (9.0)	8020 (15.9)	9273 (18.4)	50510 (100)
Human development index	0.47 (0.11)	0.63 (0.05)	0.49 (0.08)	0.71 (0.02)	0.44 (0.10)	0.54 (0.09)	0.55 (0.12)
Age, years	42.2 (14.1)	45.8 (15.3)	42.9 (13.6)	48.0 (15.1)	43.3 (13.4)	43.6 (13.2)	44.3 (14.4)
Education, years	7.5 (5.2)	7.1 (5.1)	7.3 (5.9)	12.4 (3.4)	6.9 (4.8)	7.8 (4.3)	7.8 (5.1)
Body mass index, kg/m^2	23.4 (4.0)	25.4 (3.8)	23.9 (3.9)	25.7 (3.3)	21.2 (3.3)	22.4 (3.4)	23.7 (4.0)
Health score	88.2 (14.2)	90.9 (11.3)	91.3 (12.6)	86.4 (13.2)	87.9 (14.1)	88.7 (13.5)	89.1 (13.1)
Female							
Total, N (%)	11327 (19.2)	19789 (33.6)	2051 (3.5)	8373 (14.2)	7121 (12.1)	10277 (17.4)	58938 (100)
Human development index	0.46 (0.1)	0.63 (0.05)	0.51 (0.08)	0.71 (0.02)	0.44 (0.11)	0.54 (0.09)	0.56 (0.12)
Age, years	41.9 (14.5)	45.0 (15.2)	42.2 (13.8)	49.0 (15.3)	43.0 (13.6)	42.9 (13.3)	44.3 (14.7)
Education, years	5.7 (5.0)	6.9 (5.1)	4.8 (5.7)	12.2 (3.5)	5.3 (4.8)	7.1 (4.6)	7.2 (5.2)
Body mass index, kg/m^2	24.1 (4.9)	25.8 (4.7)	24.6 (4.5)	25.9 (4.5)	21.2 (3.7)	22.1 (3.9)	24.2 (4.8)
Health score	84.2 (15.7)	87.5 (12.7)	86.9 (15.4)	82.2 (14.9)	85.2 (15.8)	87.4 (13.9)	85.8 (14.5)

Table 3.3 Effect estimate (95% confidence interval) for human development level on individual health, World Health Survey 2002-2004 (N=109,448).

	Male ***b*** **(95% CI)**	**Female** ***b*** **(95% CI)**
Scenario 1[a]		
Total effect	1.58 (-0.61, 3.77)	2.61 (-0.09, 5.32)
Pure direct effect	1.32 (-0.87, 3.51)	2.18 (-0.52, 4.88)
Total indirect effect	0.26 (0.17, 0.35)	0.44 (0.28, 0.59)
Scenario 2[b]		
Natural indirect effect via BMI only	0.016 (-0.005, 0.037)	-0.033 (-0.077, 0.011)

[a] Examining the mean difference of health score associated with 0.1-unit increase of human development index (HDI) with reference level of median HDI of 0.572, with education being the mediator of interest.

[b] BMI path-specific effect, part of the pure direct effect in Scenario 1, was further examined.

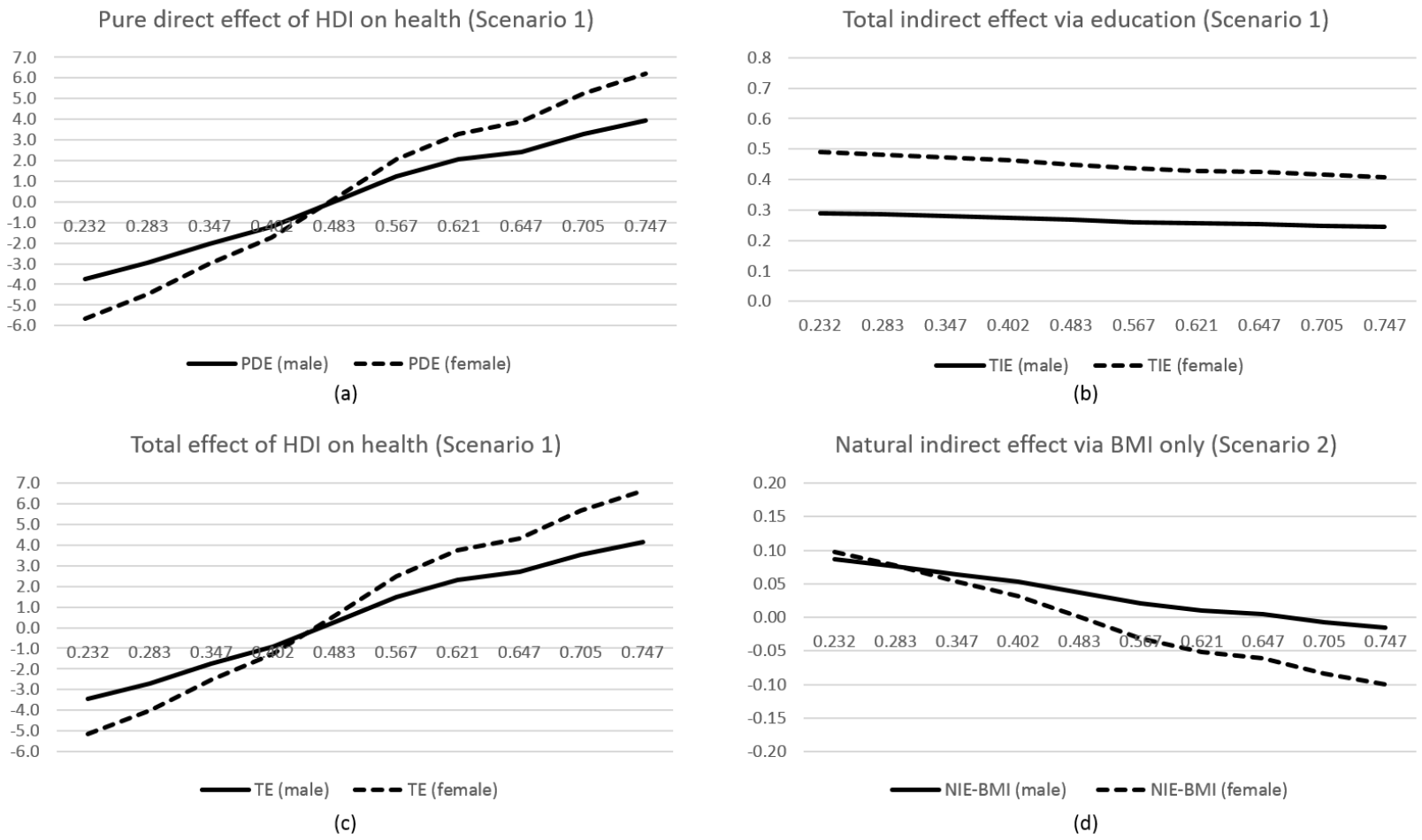

Figure 3.3 Effect decomposition when education is the mediator of interest (a-c), and natural indirect effect via BMI only in Scenario 2 (d), obtained from multilevel regression analysis of the World Health Survey 2002-2004 (N=109448). Y axis: mean difference in health score associated with a 0.1-unit increase in HDI; X axis: selected reference HDI values.

3.5 Discussion

This study examined the overall and mediated impact of HDI as a measure national human development on individual health, with possible pathways via education and BMI, using large multilevel global comparative data. We found that the HDI effect on health depended on the reference HDI level: the total effect and pure direct effect of HDI were negative at low HDI level but became positive at higher levels of HDI. Such impact was mainly through pathways other than education and BMI. The impact of HDI on health was greater for females than for males.

Our study found that at the lower end of the HDI spectrum, higher human development level predicted poorer individual health whereas at the upper end of this spectrum, higher human development level predicted better individual health. Around the median level of HDI, people from countries that were 0.1-unit apart in HDI tended to have similar health status. The HDI assesses how well countries are doing in three dimensions: health, education, and living standards.[84] The overall health status and education achievement for residents in a country will tend to grow as the country's adult literacy rate and combined gross enrollment ratio for primary, secondary and tertiary education increase. Yet, improved living standards, as captured by gross national income per capita, can affect health in a complex way. Life expectancy in LMICs has increased dramatically in the past century due to enormous achievements in control of infectious diseases via better sanitation and food safety, vaccines, antibiotics and improved nutrition.[85] On the other hand, urbanization and globalization have led to rapid changes in lifestyle such as dietary and physical activity patterns. Energy-dense, poor-quality diets and sedentary behaviors have fueled the obesity epidemic and an increased burden in nutrition-related non-communicable diseases (NCDs).[86–88] These countries were also experiencing epidemiological transition from

infectious diseases to chronic NCDs, which played an increasingly important role in people's overall health status and physical functioning, especially among the elderly. In addition, the fast economic growth in some countries may be at the expense of the environment. All these aspects contributed to the complicated relationship between development and health.

Our findings may reflect investment in and prioritization of different aspects in overall health improvement as these countries went through different stages of economic and societal development. Such investment may touch all aspects of health and health care: building roads to improve access to care, providing effective treatment for HIV/AIDS, tuberculosis, and malaria, promoting vaccination and proper use of antibiotics, strengthening primary care and preventive interventions and so on. Though all these investments and efforts may have greatly increased life expectancy at birth at the country level, they may differentially impact individual health experiences that may be heavily dependent on the presence of chronic diseases and disabilities. There might be lessons learn from countries at the lower end of the human development spectrum. All else equal, people from these countries achieved same or even higher health status compared to people from countries with median health development level, possibly through mechanisms such as targeted health investments; they may have been started experiencing the nutrition transition when the study was conducted.

Past literature has documented the presence of effect modification of HDI on the relationship between education and health,[75] and of education on the relationship between obesity and health outcomes.[73,89] Incorporating such interactions is crucial for the present study. Despite the small effect size, we found a consistent positive indirect effect of HDI on health via increased

education. In countries with lower human development levels, the education channel offset part of the negative impact of increase development on health; in countries with higher human development levels, education contributed to the overall positive impact of HDI on health. However, pathways through BMI but not education did not appear to play an important role in transmitting the impact of HDI to health. Possible explanations included that the impact of HDI on health via BMI also went through education and other upstream variables, or that the mediating role of BMI depended on HDI and/or education in a complex way that the current model did not capture. Nonetheless, pathways via education and BMI only accounted for a relatively small portion of the HDI impact. Future studies can explore pathways through other factors such as neighborhood environment, health care infrastructure, and access to care.[90]

Though women suffer more from ill health than men do;[82,91] our study suggested that women could potentially have more gains in health as a country achieved higher human development. At a low human development level, country's development had a more negative impact for women than for men. It is still unclear the underlying mechanisms for such sex difference. Women are found to be at significantly higher risk of depression.[92] In the current study sample, depression status is highly predictive of poor health.[82] It may be that the level of human development correlated with mental health services and women from countries with good mental health care (usually countries with higher human development) gain more in terms of health compared to their male counterparts.

This is the first study to examine the pathways through which country's development can affect individual's health using causal mediation analysis. The standardized methodology used in WHS

allowed for pooled analyses from middle- and low-income countries across continents. The use of causal mediation analysis enable us to incorporate nonlinear relationships such as previously documented interactions between HDI and education[75], and between education and BMI[73] in affecting individual's health outcomes. We conducted sensitivity analysis to test the robustness of the BMI-path-specific effect against the presence of confounders of the BMI-health relationship affected by HDI and/or education. We used linear mixed models to account for the multilevel structure of the data and adjusted for both contextual and individual confounders.

Several methodological limitations need to be addressed. We imposed temporality assumptions on the cross-sectionally collected WHS data: although measured at the same time, educational attainment was taken to precede current BMI, which in turn preceded current health status at the time of survey. This is a reasonable assumption in an adult population. However, there is still a slight chance that middle-age health status affected the cumulative education years. BMI tends to be stable at middle age but we cannot rule out the possibility of reverse causation for BMI-health relationship. Despite conducting sensitivity analyses to account for intermediate confounding, our result for BMI-path-specific effect of HDI could still be subject to uncontrolled confounding between BMI and health. There could be measurement error in BMI that was created based on self-report height and weight.

Our study went beyond linking contextual environment to country level health indicators such as mortality and provided evidence on the pathways through which country development could impact individual health. We also found that such impact was mainly through pathways other than education or BMI and that the impact differed by sex. Country development may harm or

benefit human health, which will eventually affect subsequent human development over time. Characterizing the impact of human development on health provides a promising initial foray into revealing and explaining the complex pathways between health and development. This can shed light on how to translate economic growth into improved health for all.

3.6 Appendix

Description of the World Health Survey (WHS) data set

Within each country, samples were probabilistically selected with every individual being assigned to a known non-zero selection probability. These samples were nationally representative except in China, Comoros, Congo, Côte d'Ivoire, India, and the Russian Federation, where the survey was carried out in geographically limited regions. This study included participants from 14 countries in the African region, nine in the European region, seven in the Americas, five in the South-East Asia region, five in the western Pacific region, and two in the Eastern Mediterranean region (Table A 3.1). All respondents were interviewed face-to-face with the standardized WHS survey, which included questions regarding demographic, socioeconomic, and behavioral factors.

Assumptions for identification

To estimate the effects defined in Table 3.1 using observational data, we assumed stable unit treatment value assumptions (SUTVA),[46,93] general consistency, conditional exchangeability (no-uncontrolled-confounding), and positivity.[53] The conditional exchangeability assumption for natural effects included: (i) no uncontrolled confounding of the (X, M) – Y and X – M relationship given covariate set **Z**, and (ii) no members of the covariate set **Z** are affected by X in scenario 1. In scenario 2, we assumed that (iii) no uncontrolled confounding of the (X, L, M) – Y, (X, L) – M, and L – M relationship given covariate set **Z**, and (iv) no members of the covariate set **Z** are affected by *X or L.* Violation of assumption (ii) in scenario 2 – confounders for the M–Y relationship being affected by either X or L or both – will be examined in sensitivity analysis.

G-estimation for sensitivity analysis

We examined how the BMI path-specific effect in scenario 2 will change under possible violation of assumption (iv) above (Figure A 3.2). We hypothesized that living in urban or rural areas, unemployment, marital status, smoking, alcohol use, and physical inactivity could be confounders of the BMI – health relationship that were influenced by HDI, education, or both and denoted these factors using **V**. In a structure presented in Figure A 3.2 (a), not adjusting for **V** will result in bias for the BMI-path-specific effect of HDI because of the extra path HDI→intermediate confounders→health. However, we cannot use traditional regression adjustment because adjusting for **V** will block the path HDI→intermediate confounders→BMI→health, which is part of the indirect effect of HDI via BMI but not education. In this case, using g-estimation to de-activate the path from intermediate confounders to health will be appropriate for unbiased estimation of this targeted effect as presented in Figure A 3.2 (b). We created the below new outcome variable Y^* where $Y^* \perp \mathbf{V} \mid X, L, M, \mathbf{Z}$:

$$Y^* = Y - \boldsymbol{v}\varphi_V + \varphi_V \cdot E(\mathbf{V})$$

where φ_V came from the model for Y with additional adjustment for **V**, i.e., $E(Y|x, m, \boldsymbol{z}, \boldsymbol{v}; \varphi) = \varphi_Y + \varphi_X \cdot x + \varphi_{X^2} \cdot x^2 + \varphi_L \cdot l + \varphi_M \cdot m + \varphi_{XL} \cdot x \cdot l + \varphi_{XM} \cdot x \cdot m + \varphi_{LM} \cdot l \cdot m + \boldsymbol{z}\varphi_Z + \boldsymbol{v}\varphi_V$**,** and $E(\mathbf{V})$ represented the crude expected value for **V**. Then, we used this new Y^* variable for all analyses.

Appendix Tables

Table A 3.1 Country-specific sample size, percent female, mean age, and national human development index, World Health Surveys 2002-2004.

Country	Initial sample size	Human development index 1990	N missing health score	N missing education	Final N (appendix result, scenario 1)	N missing height or weight	Final N (main result)	Female (%)	Mean age
African Region (AFR)									
Burkina Faso	3605	Missing	91	749	0	2328	0	50.7	41.4
Chad	3624	Missing	364	1177	0	748	0	51.7	41.8
Comoros	1411	Missing	55	30	0	14	0	57.0	47.5
Congo	1935	0.553	413	1170	673	225	651	52.4	40.0
Côte d'Ivoire	2398	0.380	248	928	1353	222	1255	42.0	40.3
Ethiopia	3772	Missing	442	2381	0	3138	0	51.1	41.9
Ghana	3292	0.502	113	339	2855	214	2609	55.6	45.1
Kenya	3441	0.471	46	256	3144	295	2842	57.9	42.6
Malawi	3690	0.283	124	371	3203	213	2956	56.5	42.3
Mali	3095	0.232	2389	2623	90	1056	35	43.2	46.2
Mauritania	3008	0.367	294	854	1942	437	1558	61.7	43.2
Mauritius	3385	0.621	302	42	3045	1239	1867	52.7	45.2
Namibia	3283	0.577	1200	352	1947	321	1782	59.3	42.6
Senegal	2527	0.384	548	1381	956	1110	590	48.2	42.9
South Africa	1869	0.619	330	0	1539	624	866	53.2	41.8
Swaziland	2364	0.538	851	748	1307	940	1002	53.8	43.8
Zambia	2839	0.407	333	87	2424	1062	1502	53.5	41.2
Zimbabwe	3013	0.488	117	48	2863	1014	1685	64.9	43.1
Region of the Americas(AMR)									
Brazil	4209	0.612	548	42	3622	480	3194	56.8	45.6
Dominican Republic	3758	0.589	61	17	3685	1199	2458	53.2	45.8
Ecuador	3866	0.643	326	461	3126	444	2648	55.7	45.0

Country	Initial sample size	Human development index 1990	N missing health score	N missing education	Final N (appendix result, scenario 1)	N missing height or weight	Final N (main result)	Female (%)	Mean age
Guatemala	3822	0.483	143	754	2955	1140	2096	61.1	44.6
Mexico	32129	0.647	0	0	32129	12689	19272	57.5	45.1
Paraguay	4062	0.581	45	1	4017	357	3597	54.6	44.9
Uruguay	2680	0.691	22	5	2654	2	2596	51.8	48.7
Eastern Mediterranean Region (EMR)									
Morocco	4184	0.459	4184	2473	0	2538	0	58.3	44.9
Pakistan	5027	0.402	190	696	4178	2400	1957	45.3	41.6
Tunisia	4213	0.567	344	527	3411	737	2804	54.9	45.9
European Region (EUR)									
Bosnia and Herzegovina	917	Missing	386	1	0	3	0	58.3	50.1
Croatia	932	0.689	20	4	909	8	885	59.9	54.1
Czech Republic	828	0.762	90	16	724	18	688	55.6	51.3
Estonia	927	0.73	41	3	884	5	849	63.8	52.3
Georgia	2441	Missing	16	8	0	7	0	57.9	52.2
Hungary	1262	0.701	315	0	947	17	906	59.4	53.0
Kazakhstan	4110	0.686	104	9	3997	330	3621	65.8	43.3
Latvia	763	0.71	72	18	679	116	568	68.3	54.6
Russian Federation	4068	0.729	278	174	3629	869	2838	64.6	54.0
Slovakia	1917	0.747	680	585	1222	585	1185	63.8	43.8
Ukraine	2517	0.705	205	37	2275	920	1395	65.3	50.8
South-East Asia Region (SEAR)									
Bangladesh	4526	0.382	821	1380	2633	3891	500	52.2	42.6
India	8139	0.431	1640	1371	5492	1911	4196	51.7	43.0
Myanmar	4996	0.347	4	0	4992	30	4946	57.3	44.6
Nepal	6979	0.388	49	426	6511	4694	2146	56.3	43.3
Sri Lanka	5642	0.62	710	673	4409	1328	3353	54.0	44.9

Country	Initial sample size	Human development index 1990	N missing health score	N missing education	Final N (appendix result, scenario 1)	N missing height or weight	Final N (main result)	Female (%)	Mean age
Western Pacific Region (WPR)									
China	3674	0.502	54	17	3603	4	3579	51.4	47.2
Lao People's Democratic Republic	4060	0.395	86	5	3969	9	3923	52.7	41.8
Malaysia	5249	0.641	203	153	4909	999	3910	56.8	44.2
Philippines	8378	0.591	110	24	8245	1582	6614	54.6	42.6
Viet Nam	2982	0.476	1427	35	1532	7	1524	55.5	43.4

Table A 3.2 Participant characteristics by WHO region, World Health Survey 2002-2004 (N=148,679).

Characteristics, mean (SD)	Africa	The Americas	Eastern Mediterranean	Europe	South-East Asia	Western Pacific	All
Male							
Total, N (%)	12216 (18.6)	22724 (34.6)	3926 (6.0)	5419 (8.2)	11250 (17.1)	10201 (15.5)	65736 (100)
Human development index	0.47 (0.11)	0.63 (0.04)	0.47 (0.08)	0.71 (0.02)	0.43 (0.09)	0.54 (0.09)	0.55 (0.12)
Age, years	42.7 (14.4)	45.7 (15.4)	43.4 (14.2)	48.7 (15.3)	44.3 (14.0)	43.9 (13.4)	44.7 (14.7)
Education, years	7.3 (5.1)	7.1 (5.1)	6.6 (5.9)	12.3 (3.5)	6.0 (4.9)	7.7 (4.3)	7.4 (5.1)
Health score	87.7 (14.8)	90.8 (11.5)	90.2 (13.5)	85.9 (13.8)	86.0 (15.9)	88.2 (13.9)	88.6 (13.8)
Female							
Total, N (%)	15125 (18.2)	29464 (35.5)	3663 (4.4)	9847 (11.9)	12787 (15.4)	12057 (14.5)	82943 (100)
Human development index	0.47 (0.10)	0.63 (0.04)	0.48 (0.08)	0.71 (0.02)	0.43 (0.09)	0.55 (0.09)	0.56 (0.12)
Age, years	42.5 (14.6)	44.9 (15.2)	42.5 (14.2)	50 (15.7)	42.8 (13.8)	43.5 (13.7)	44.4 (14.9)
Education, years	5.7 (4.9)	6.6 (5.0)	3.5 (5.1)	12.0 (3.6)	4.1 (4.6)	7.0 (4.5)	6.6 (5.2)
Health score	83.8 (16.1)	87.5 (12.9)	83.5 (17.0)	81.3 (15.6)	82.0 (18.0)	86.8 (14.2)	84.9 (15.3)

Table A 3.3 Effect estimate (95% confidence interval) for scenario 1[a], World Health Survey 2002-2004 (N=148,679).

	Male	Female
Total effect	1.54 (-0.74, 3.83)	2.42 (0.44, 4.41)
Pure direct effect	1.29 (-0.99, 3.57)	2.00 (0.02, 3.98)
Total indirect effect	0.26 (0.17, 0.34)	0.42 (0.30, 0.55)

[a] Examining the mean difference of health score associated with 0.1-unit increase of human development index (HDI) with reference level of median HDI of 0.572, with education being the mediator of interest.

Table A 3.4 Potential confounders for BMI-health relationa by WHO region, World Health Survey 2002-2004 (N=105,630).

Characteristics, mean (SD)	Africa	The Americas	Eastern Mediterranean	Europe	South-East Asia	Western Pacific	All
Living in rural areas, N (%)	13240 (61.55)	11093 (31.66)	1815 (42.14)	3398 (27.38)	11313 (75.19)	10513 (53.22)	51372 (47.54)
Unemployment, N (%)	8710 (40.49)	15602 (44.52)	1993 (46.27)	4905 (39.52)	4951 (32.91)	6390 (32.35)	42551 (39.37)
Not Married, N (%)	7635 (35.5)	15485 (44.19)	866 (20.11)	4809 (38.75)	3037 (20.18)	3553 (17.99)	35385 (32.74)
Currently smoking, N (%)	3616 (16.81)	8031 (22.92)	1215 (28.21)	3396 (27.36)	5418 (36.01)	6619 (33.51)	28295 (26.18)
Alcohol use, N (%)	6926 (32.2)	19785 (56.46)	308 (7.15)	9700 (78.16)	2900 (19.27)	7433 (37.63)	47052 (43.54)
Physical inactivity, N (%)	13663 (63.52)	28328 (80.84)	3379 (78.45)	9061 (73.01)	8630 (57.36)	12858 (65.1)	75919 (70.25)

[a] Factors examined in sensitivity analyses and their statistics were from sample further restricted to individuals have complete information on these factors.

Appendix Figures

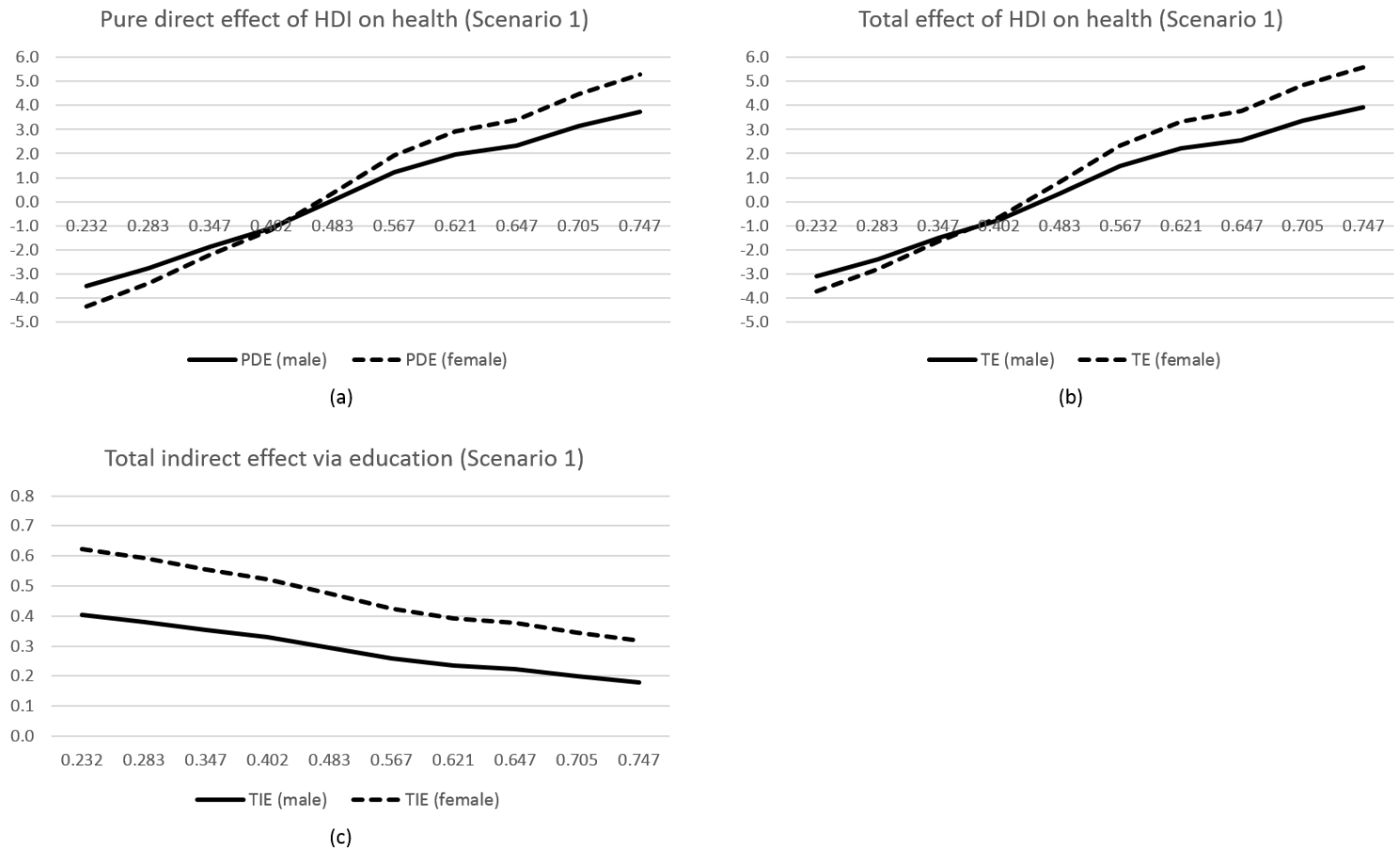

Figure A 3.1 Effect decomposition when education is the mediator of interest (a-c), using the larger sample of 148,679 participants, the World Health Survey 2002-2004. Y axis: mean difference in health score associated 0.1-unit increase in HDI; X axis: selected HDI values within the range of the current sample.

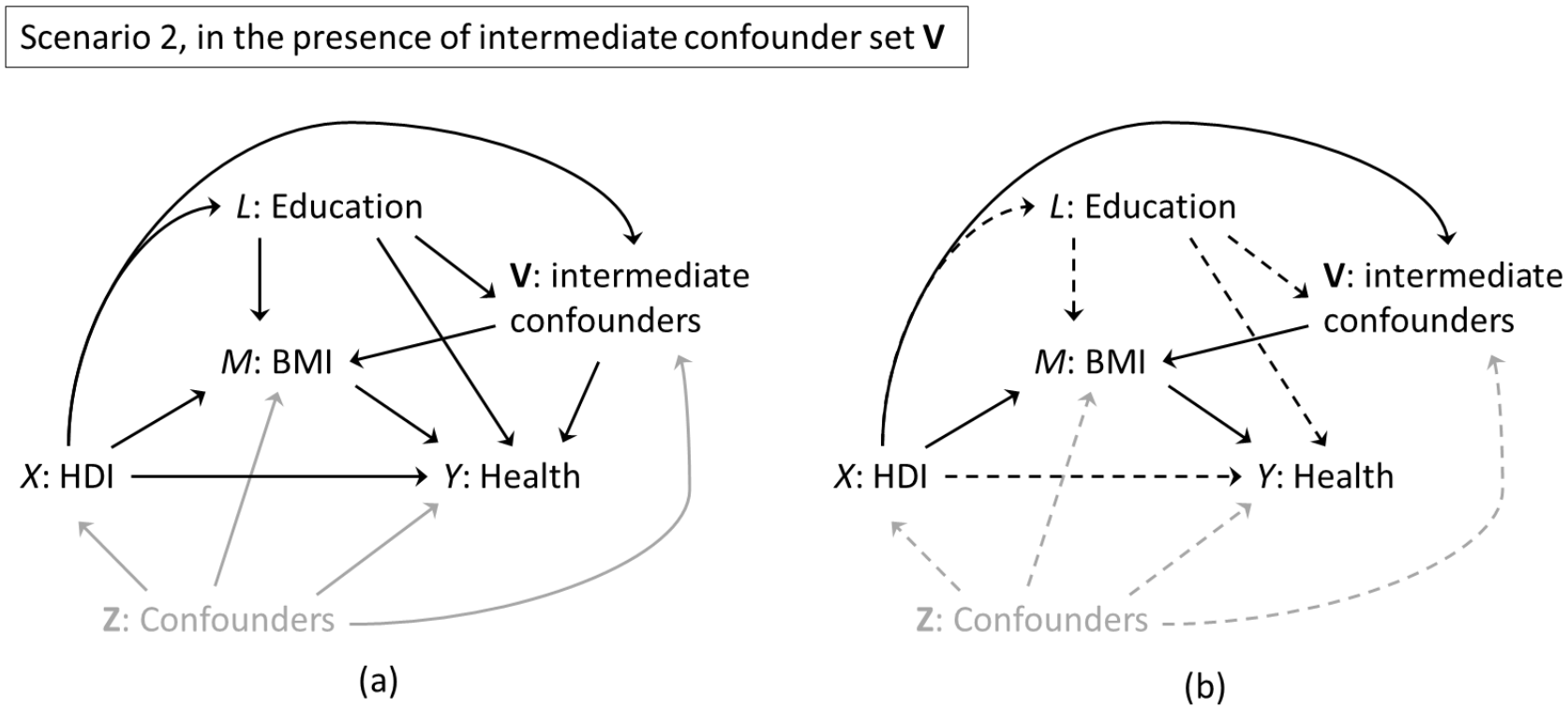

Figure A 3.2 Graphical representation of scenario 2 in the presence of *M*-*Y* confounder set **V** that was affected by *X* and *L* (a) and the unbiased indirect effect of HDI on health via BMI only (solid line) after implementing g-estimation (b).

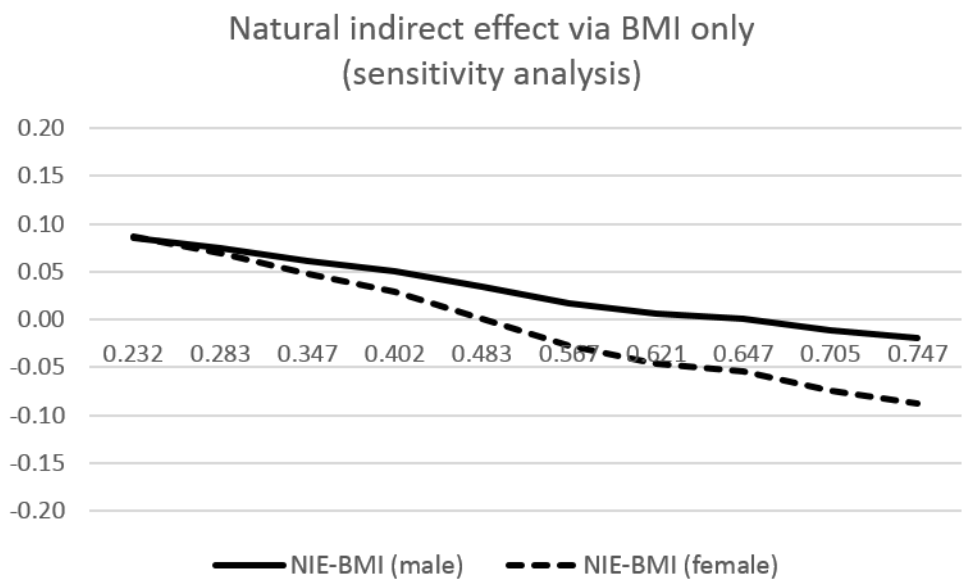

Figure A 3.3 Results from sensitivity analyses for natural indirect effect via BMI only in Scenario 2, the World Health Survey 2002-2004 (N=109,448). Y axis: mean difference in health score associated 0.1-unit increase in HDI; X axis: selected HDI values within the range of the current sample.

Chapter 4. Education and Health in 46 Countries: Modeling the Mediating Role of Social Factors and Health Behaviors

4.1 Abstract

Introduction: Past studies examining the pathways from socioeconomic status to health via health behaviors seldom considered multiple mediators and rarely accounted for the possible interactions between exposure and downstream behaviors. Using causal mediation analysis, this study examined the health disparities in education and the contributions of pathways through social mediating factors and health behaviors.

Methods: We analyzed data on 164,743 participants aged 25 or older from 46 countries across continents, collected by the World Health Survey 2002-2004. The outcome was a health score based on measures from eight health state domains. Exposure was individual educational attainment (no formal education, less than primary school, primary school completed, secondary school completed, and high school or beyond). The social mediating factors were residence, unemployment, and marriage. The mediating health behaviors were smoking, alcohol use, physical inactivity, and stress. G-computation algorithm implemented using Monte Carlo simulation of generalized linear mixed models was used to estimate natural and controlled direct effects, and pathway effects, comparing 'lower education' to 'high school or beyond' (reference) education.

Results: Lower educational attainment had an overall negative impact on health (regression coefficient *b* ranging from -1.06 for secondary school completed to -4.05 for no formal education), the largest proportion of which was neither mediated by social factors nor health behaviors. Pathways through only health behaviors accounted for around one tenth of the total effect of education. A substantial amount of the observed health disparities would be eliminated

if everyone had healthy behaviors (proportion eliminated ranging from 48% for secondary school completed to 72% for no formal education).

Conclusion: Simultaneous intervention on education together with health behaviors or social factors will be more effective in reducing health disparity than intervention on education alone, since these subsequent mediating factors were important effect modifiers for the education-health relation.

4.2 Introduction

The literature has persistently documented the relationship between lower socioeconomic status (SES) and poor health among affluent countries[94–98] and worldwide.[82] However, the mechanisms underlying education and health are not conclusive. Though different factors and pathways have been proposed to explain the socioeconomic inequalities in health,[90,99–104] health behavior/lifestyle appeared to be the most popular factor. It is suggested that unhealthy behavior or lifestyle accounted for 50% of mortality in 1976 in the U.S., according to the Healthy People: The Surgeon-General's Report on Health Promotion and Disease Prevention in 1979[105], highlighting the importance of studying these behavioral factors in explain the educational gradient in health.

Empirical evidence is abundant on the relations between low SES and health behaviors such as smoking, alcohol consumption, physical inactivity,[106–108] or psychosocial factors,[109,110] and between these behaviors and various health comes.[100,107,111,112] Some studies also examined the pathways or the direct impact of SES on health accounting for possible mediating pathways.[96,99,106,107,110,113,114] However, these studies were mainly from European countries and

the U.S., results reflecting a global picture is lacking. Also, past studies using traditional regression approaches such as the "difference method"[14] or linear path model[6] seldom consider the possible interactions between SES and downstream behaviors. Past decade has seen blooming literatures on the effect definition, identification, and estimation issues for causal mediation analysis under the potential outcome framework.[8,11,12,19,34,35] Such work also extends to complex settings involving multiple causally ordered mediators[28,52,53] and time-varying exposure and mediators.[32]

In this study, we examined (1) health disparities due to educational attainment, a commonly used SES measure that shapes future occupational opportunities and income,[100] and (2) the contributions of possible mediating pathways through social factors and health behaviors. We partitioned the total effect of education on health into pathway effects while preserving the interaction between education and mediators (mechanistic perspective). We also examine two types of controlled direct effect that correspond to certain hypothetical public health interventions on social factors and health behaviors.

4.3 Method

Study sample and variables

We used data from the cross-sectional World Health Survey (WHS) conducted by the WHO in 70 high-, middle-, and low-income countries from 2002 to 2004. The study design and methods of the WHS have been documented in detail in the appendices and elsewhere.[81,82] The standardized methodology allowed for examination of individual health measures across countries.[81]

Outcome

The health state measures has been extensively tested[83] and have showed good consistency and reliability.[82] Individual participants were asked to report their perceived difficulties based on a 5-point Likert scale question for eight health state domain. These domains are mobility, self-care, pain and discomfort, cognition, interpersonal activities, vision, sleep and energy, and affect, each of which consists of two questions.[81] We performed factor analysis using polychoric correlations to account for the covariance structure of the responses to individual questions. Similar to a previous study,[82] we chose one factor solution based on the high eigenvalue of the first factor (8.92, 73% as a cumulative percentage of the variance explained) and the high communalities of the original variables (between 0.36 and 0.70). Then, we used the principal component method for factor extraction and the regression scoring method to obtain the factor scores. The factor score was rescaled with 0 indicating worst health and 100 indicating best health.

Exposure

Individual educational attainment was measured as the highest level of education a person completed. There are seven categories: “no formal schooling”, “less than primary school”, “primary school completed”, “secondary school completed”, “high school (or equivalent) completed”, “college/preuniversity/university completed” or “post graduate degree completed”. The last three categories are combined into “high school or beyond”. When an individual’s educational attainment was missing but reported having 0 years of schooling, we assigned “no formal schooling” as their educational attainment (N=37). These categories were made to be

applicable to all countries regardless of the type of educational system via a mapping algorithm to record educational categories other than specified above.[63]

Mediators

Social factors included residence (living in rural areas versus urban or semi-urban area), unemployment (currently not employed versus employed), and marital status (currently not married versus married).

*Individual health behavior*s included smoking (currently smoke versus not), alcohol drinking (ever versus never), physical inactivity (having <3 times of vigorous physical activity per week versus having ≥3 times) and stress. Participants were asked "How often have you felt that you were unable to control the important things in your life" and "How often have you found that you could not cope with all the things that you had to do". Answers based on 5-point Likert scale (ranging from "1-never" to "5-very often") were aggregated and then log transformed (using log base 2) and re-centered so that higher scores indicated more stress while 0 represented no stress (score ranging from 0 to 2.3).

Confounders

Potential contextual confounder are WHO region and country level wealth, measured by gross domestic product per capita (in current US$) in 2003 that ware obtained from the UN database.[115] Individual level predisposing factors are age and sex.

Conceptual framework

We used a directed acyclic graph (DAG)[24] to represent our assumptions about the data generating process (Figure 4.1). Let *Y* denote one's health state, *X* the educational attainment,

$\mathbf{M_A}$ the social factors – the first set of mediators of interest, $\mathbf{M_B}$ the individual health behaviors – the second set of mediators of interest, and **Z** the set of covariates not affected by the exposure but which are assumed to be sufficient for confounding control for effects estimation. Let $Y(x, \mathbf{M_A}(x), \mathbf{M_B}(x, \mathbf{M_A}(x)))$ represent the potential Y had X been set to x, $\mathbf{M_A}$ been set to the natural value under $X=x$, $\mathbf{M_B}$ been set to the natural value under $X=x$ and $\mathbf{M_A} = \mathbf{M_A}(x)$. Let x denote any one of the four index education level: "no formal education", "less than primary school", "primary school completed", and "secondary school completed", and x^* the reference level: "high school or beyond", which were the two values of the exposure we wish to compare. Let $\mathbf{m_a}^*$ and $\mathbf{m_b}^*$ denote the reference values for $\mathbf{M_A}$ and $\mathbf{M_B}$ used in controlled direct effects. We consider factors in $\mathbf{M_A}$ and $\mathbf{M_B}$ jointly as a construct without further specifying the causal direction between factors in the same set.

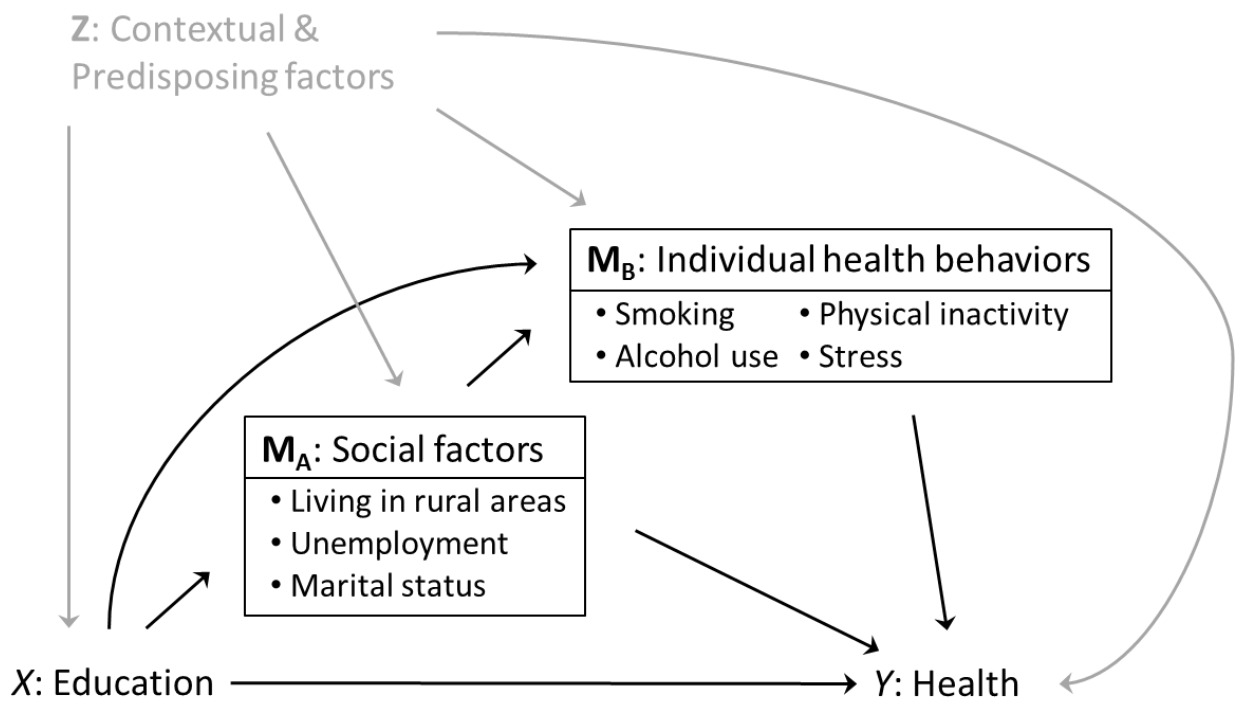

Figure 4.1 DAG depicting a single exposure X, two sets of causally ordered mediators MA and MB, an outcome Y, and a set of confounders Z sufficient for confounding control.
DAG depicting a single exposure *X*, two sets of causally ordered mediators $\mathbf{M_A}$ and $\mathbf{M_B}$, an outcome Y, and a set of confounders **Z** sufficient for confounding control. When only mediator $\mathbf{M_B}$ is of interest, $\mathbf{M_A}$ is sometimes called "intermediate confounders" or "endogenous confounders" as they are consequences of *X*.

Effect decomposition, definition, and empirical analogs

In the presence of two causally ordered mediators, total effect (TE) of education on health can be decomposed into a natural direct effect (NDE) of education that is not through any of the two mediator sets, a natural indirect effect through social factors and their consequences (NIE-A), and a natural indirect effect through health behaviors only (NIE-B) (Figure 4.2). We also examined two types of controlled direct effect: (i) one that captured the direct effect of education had we fixed both social factors and health behaviors at their reference levels (CDE-00), and (ii) one that captured the direct effect of education while fixing only health behaviors at the

reference levels but not fixing social factors (i.e., allowing them to respond to education) (CDE-X0). Effect definitions under the potential outcome (counterfactual) framework were listed in Table 4.1. There are other ways to decompose the total effect into components representing natural direct and indirect effects.[53] The current decomposition was also discussed in the intermediate-confounding context[28,116] and has the advantage of circumventing the need for specifying an additional sensitivity parameter. This parameter represents the conditional correlation between $\mathbf{M_A}(x)$ and $\mathbf{M_A}(x^*)$ given $\mathbf{Z}$ and cannot be obtained from the observed data. Notice that each potential outcome expression (half of the effect definition) listed in Table 4.1 (left column) follows the form of $Y(x_1, \mathbf{M_A}(x_2), \mathbf{M_B}(x_3, \mathbf{M_A}(x_4)))$ and for all listed expressions, we have $x_2 = x_4$ and thus is a special case.[53] We invoked the stable unit treatment value assumption (SUTVA) [46,93], and assumptions of general consistency, positivity, and conditional exchangeability (no uncontrolled confounding).[53] Details for conditional exchangeability assumption are listed in the appendix.

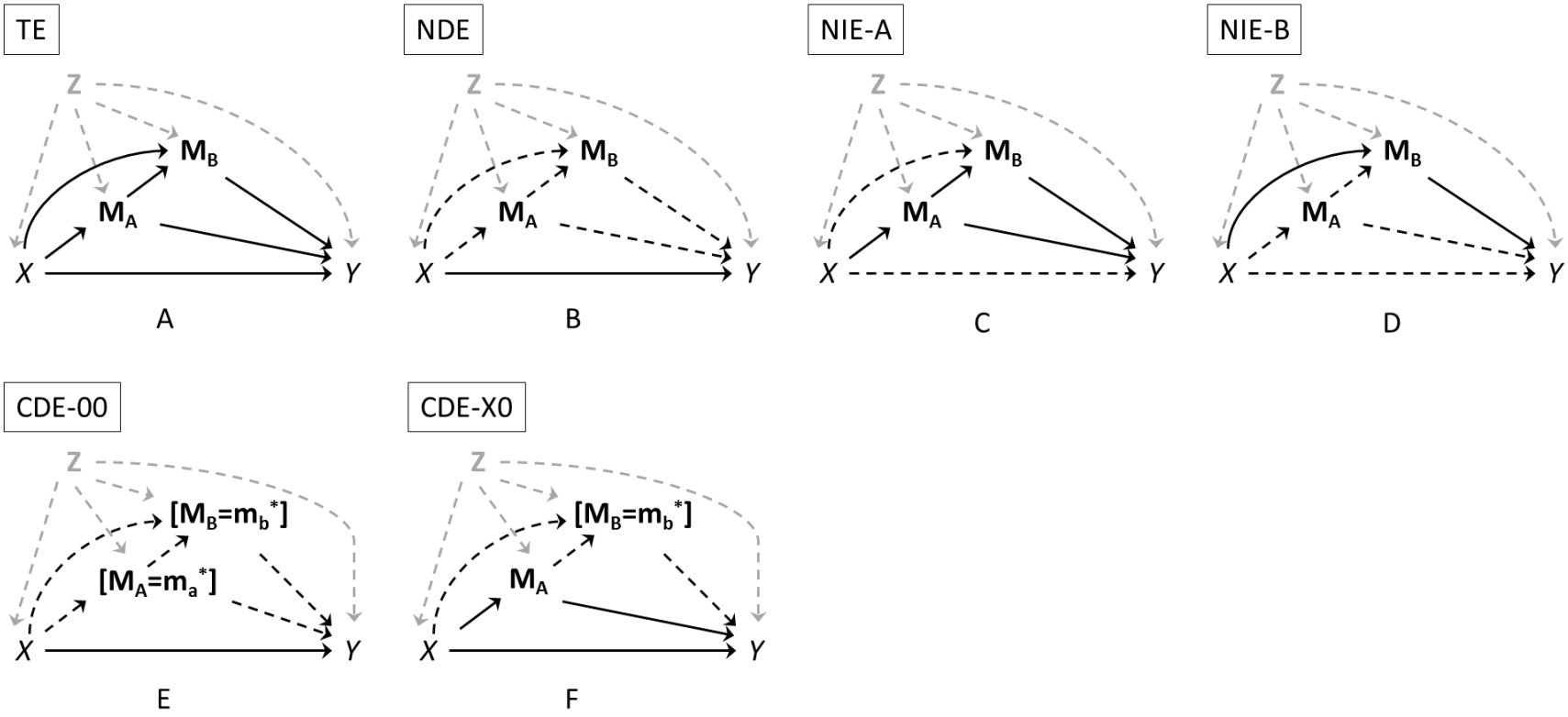

Figure 4.2 Graphical representations for different types of effect of interest. *X*: educational attainment, $\mathbf{M_A}$: social factors including residence, unemployment, and marital status; $\mathbf{M_B}$: individual health behaviors including smoking, alcohol use, physical inactivity, and stress; *Y*: individual's health; and **Z**: contextual and predisposing factors including WHO region, country level wealth and individual's age and sex; TE: total effect; NDE: natural direct effect; NIE-A: natural indirect effect that is through $\mathbf{M_A}$ and its consequences; NIE-B: natural indirect effect that is through $\mathbf{M_B}$ only; CDE-00: controlled direct effect while fixing $\mathbf{M_A}=\mathbf{m_a}^*$ and $\mathbf{M_B}=\mathbf{m_b}^*$; CDE-X0: controlled direct effect while fixing only $\mathbf{M_B}=\mathbf{m_b}$ and allowing $\mathbf{M_A}$ to respond to *X*. Black solid lines represent the corresponding effect of *X* on *Y*. NDE, NIE-A and NIE-B add up to TE.

Table 4.1 Effect definition and empirical analogs, applied to World Health Survey 2002-2004[a].

Effect	Counterfactual definition	Empirical analog[b]
TE	$\mathrm{E}\{Y(x, \mathbf{M_A}(x), \mathbf{M_B}(x, \mathbf{M_A}(x))) - Y(x^*, \mathbf{M_A}(x^*), \mathbf{M_B}(x^*, \mathbf{M_A}(x^*)))\}$	$\sum_{\mathbf{z}}\sum_{\mathbf{m_a}}\sum_{\mathbf{m_b}}\{E(Y\|x,\mathbf{m_a},\mathbf{m_b},\mathbf{z})P(\mathbf{m_b}\|x,\mathbf{m_a},\mathbf{z})P(\mathbf{m_a}\|x,\mathbf{z}) - E(Y\|x^*,\mathbf{m_a},\mathbf{m_b},\mathbf{z})P(\mathbf{m_b}\|x^*,\mathbf{m_a},\mathbf{z})P(\mathbf{m_a}\|x^*,\mathbf{z})\}P(\mathbf{z})$
NDE	$\mathrm{E}\{Y(x, \mathbf{M_A}(x^*), \mathbf{M_B}(x^*, \mathbf{M_A}(x^*))) - Y(x^*, \mathbf{M_A}(x^*), \mathbf{M_B}(x^*, \mathbf{M_A}(x^*)))\}$	$\sum_{\mathbf{z}}\sum_{\mathbf{m_a}}\sum_{\mathbf{m_b}}\{E(Y\|x,\mathbf{m_a},\mathbf{m_b},\mathbf{z}) - E(Y\|x^*,\mathbf{m_a},\mathbf{m_b},\mathbf{z})\}P(\mathbf{m_b}\|x^*,\mathbf{m_a},\mathbf{z})P(\mathbf{m_a}\|x^*,\mathbf{z})P(\mathbf{z})$
NIE-A	$\mathrm{E}\{Y(x, \mathbf{M_A}(x), \mathbf{M_B}(x^*, \mathbf{M_A}(x))) - Y(x, \mathbf{M_A}(x^*), \mathbf{M_B}(x^*, \mathbf{M_A}(x^*)))\}$	$\sum_{\mathbf{z}}\sum_{\mathbf{m_a}}\sum_{\mathbf{m_b}}E(Y\|x,\mathbf{m_a},\mathbf{m_b},\mathbf{z})P(\mathbf{m_b}\|x^*,\mathbf{m_a},\mathbf{z})\{P(\mathbf{m_a}\|x,\mathbf{z}) - P(\mathbf{m_a}\|x^*,\mathbf{z})\}P(\mathbf{z})$
NIE-B	$\mathrm{E}\{Y(x, \mathbf{M_A}(x), \mathbf{M_B}(x, \mathbf{M_A}(x))) - Y(x, \mathbf{M_A}(x), \mathbf{M_B}(x^*, \mathbf{M_A}(x)))\}$	$\sum_{\mathbf{z}}\sum_{\mathbf{m_a}}\sum_{\mathbf{m_b}}E(Y\|x,\mathbf{m_a},\mathbf{m_b},\mathbf{z})\{P(\mathbf{m_b}\|x,\mathbf{m_a},\mathbf{z}) - P(\mathbf{m_b}\|x^*,\mathbf{m_a},\mathbf{z})\}P(\mathbf{m_a}\|x,\mathbf{z})P(\mathbf{z})$
CDE-00[c]	$\mathrm{E}\{Y(x, \mathrm{M_A}=\mathbf{m_a}^*, \mathrm{M_B}=\mathbf{m_b}^*) - Y(x^*, \mathrm{M_A}=\mathbf{m_a}^*, \mathrm{M_B}=\mathbf{m_b}^*)\}$	$\sum_{\mathbf{z}}\{E(Y\|x,\mathbf{m_a}^*,\mathbf{m_b}^*,\mathbf{z}) - E(Y\|x^*,\mathbf{m_a}^*,\mathbf{m_b}^*,\mathbf{z})\}P(\mathbf{z})$
CDE-X0[d]	$\mathrm{E}\{Y(x, \mathrm{M_A}(x), \mathrm{M_B}=\mathbf{m_b}^*) - Y(x^*, \mathbf{M_A}(x^*), \mathbf{M_B}=\mathbf{m_b}^*)\}$	$\sum_{\mathbf{z}}\sum_{\mathbf{m_a}}\{E(Y\|x,\mathbf{m_a},\mathbf{m_b}^*,\mathbf{z})P(\mathbf{m_a}\|x,\mathbf{z}) - E(Y\|x^*,\mathbf{m_a},\mathbf{m_b}^*,\mathbf{z})P(\mathbf{m_a}\|x^*,\mathbf{z})\}P(\mathbf{z})$

[a] Y: health score, X: educational attainment (x represents each index level of education and x^* represents the reference level of education – high school or beyond), $\mathbf{M_A}$: social factors including residence, unemployment, and being unmarried, $\mathbf{M_B}$: individual health behaviors including smoking, alcohol use, physical inactivity, and stress), $\boldsymbol{Z}$: age, sex, country level gross domestic product per capita (in current US$) in 2003, and WHO region.

[b] We use $E(Y|x, \mathbf{m_a}, \mathbf{m_b}, \mathbf{z})$ as a shorthand for $E(Y|X = x, \mathbf{M_A} = \mathbf{m_a}, \mathbf{M_B} = \mathbf{m_b}, \mathbf{Z} = \mathbf{z})$, $P(\mathbf{m_b}|x, \mathbf{m_a}, \mathbf{z})$ as a shorthand for $P(\mathbf{M_B} = \mathbf{m_b}|X = x, \mathbf{M_A} = \mathbf{m_a}, \mathbf{Z} = \mathbf{z})$ and $P(\mathbf{m_a}|x, \mathbf{z})$ as a shorthand for $P(\mathbf{M_A} = \mathbf{m_a}|X = x, \mathbf{Z} = \mathbf{z})$. Summations are replaced by integrals and the probability functions by appropriate density functions for continuous variables (e.g. stress).

[c] CDE-00 represents the controlled direct effect of education on health when participants lived in urban area, being employed and married, did not smoke nor drink alcohol, being physically active and having no stress ($\mathbf{m_a}^*$ and $\mathbf{m_b}^*$ equal to zero).

[d] CDE-X0 represents the controlled direct effect of education on health when participants did not smoke nor drink alcohol, being physically active and having no stress ($\mathbf{m_b}^*$ equals to zero).

Statistical analysis

We used appropriate descriptive statistics to summarize the characteristics of the participants by their educational attainment. Under the assumptions mentioned above, each effect can be expressed in terms of their empirical analogs (Table 4.1, right column). For a specific effect, each half of the empirical analog used to estimate the expected potential outcome under different exposure and mediator assignment is recognized as an extension of the g-computation formula[39] or the mediation formula[19] to multiple-mediator settings. We adopted a fully parametric approach, implemented via Monte Carlo simulation, to obtain marginal estimates for each effect. Detailed steps can be found in Appendices. Briefly, we estimated parameters for predicting each mediator and the outcome using multilevel generalized linear models with random intercept for country. Then we created an education intervention variable, simulated the potential mediators and outcomes sequentially based on the counterfactual definition of each effect. Finally, we ran a marginal structural model to obtain a marginal estimate for each type of effect and used non-parametric bootstrap to obtain standard errors and 95% confidence intervals. All analyses were conducted in SAS 9.4 (SAS Institute Inc., Cary, North Carolina, USA).

4.4 Result

Among 231,274 participants aged 25 years or older, 164,743 (71.2%) participants from 46 countries had complete information on all covariates. Country-specific sample size

and characteristics, and the characteristics of the excluded participants are presented in appendix (Tables A 4.1 and A 4.2 respectively). The main reason for exclusion was missing information on at least one health behavior factor or health score. Excluded participants were slightly older, more likely to obtain education greater than high school, and be unemployed, unmarried, and physically inactive but less likely to have used alcohol. Distributions of other variables were similar to that among participants in the analytic sample.

Table 4.2 shows participant characteristics by educational attainment. As education increased, participants were younger, less likely to live in rural areas or be unemployed, more likely to have used alcohol and be physically inactive, and reported better health. More females and higher level of stress were seen among people with no formal education than the rest of the participants.

Table 4.3 displays the effect estimates comparing each of the four index education levels to the reference 'high school or beyond' category. Across all education levels, low educational attainment was associated with poorer health (TE ranging from -4.05 to -1.06). The impact of low education on health was negative through pathways involving social factors (NIE-A ranging from -0.54 to -0.18), pathway through health behaviors only (NIE-B ranging from -0.43 to -0.14), and pathways other than through social factors

or health behaviors (NDE ranging from -3.08 to -0.74). Lower educational attainment was associated with poorer health when we either fixed all mediators at the desired reference levels (i.e. living in urban areas, being employed and married, not currently smoking, never used alcohol, being physically active, and not feeling stressed) (CDE-00 ranging from -0.22 to -0.34), or fixed only health behaviors at the desired levels (CDE-X0 ranging from -1.12 to -0.55). One exception is that, after fixing all mediators at the aforementioned desired levels, 'no formal education' was not associated with health (CDE-00: -0.10, 95% CI: -0.40-0.19). For all types of effects, effect sizes became smaller as education level increased from "no formal education" to "secondary school completed".

Figure 4.3 depicts the proportion explained by each type of effect relative to the TE across different education levels. Indirect effect of education through social factors and their consequences made up 13.4% ~ 16.9% of the TE. Pathway involving health behaviors only accounted for around one tenth (10.6% ~ 13.3%) of the TE. The negative impact of lower education on health was mainly direct (i.e., via other pathways) (69.8% ~ 76.0%). The majority of the negative impact of lower educational attainment on health could be prevented if we could, by some hypothetical intervention, fix both social factors and health behaviors at the desired levels, especially for those with no formal education [PE(CDE-00): 97.5%, PE(CDE-00) ranging from 67.4% to 82.1% for other education

levels]. A large portion of the health disparities due to education could be eliminated if hypothetical intervention was implemented to fix health behaviors at the desired levels [PE(CDE-X0) ranging from 47.8% to 72.3%].

Table 4.2 Participant characteristics by educational attainment, World Health Survey 2002-2004 (N=164,743).

Characteristics	No formal education	Less than primary school	Primary school completed	Secondary school completed	High school or beyond	All
Total, N (%)	35973 (21.8)	20958 (12.7)	32018 (19.4)	40663 (24.7)	35131 (21.3)	164743 (100)
Age, mean (SD)	47.9 (16.1)	46.7 (15.8)	45.6 (15.6)	43.3 (14.3)	41.8 (12.8)	44.8 (15.0)
Females, N (%)	22784 (63.3)	11107 (53.0)	16847 (52.6)	21923 (53.9)	18609 (53.0)	91270 (55.4)
Social factors						
Living in rural areas, N (%)	28093 (78.1)	14213 (67.8)	16739 (52.3)	13965 (34.3)	8833 (25.1)	81843 (49.7)
Unemployment, N (%)	17005 (47.3)	8960 (42.8)	14116 (44.1)	18653 (45.9)	10711 (30.5)	69445 (42.2)
Not Married, N (%)	10213 (28.4)	7523 (35.9)	10394 (32.5)	12962 (31.9)	11249 (32.0)	52341 (31.8)
Individual health behaviors						
Currently smoking, N (%)	9438 (26.2)	5893 (28.1)	8187 (25.6)	10297 (25.3)	9006 (25.6)	42821 (26.0)
Alcohol use, N (%)	8045 (22.4)	8201 (39.1)	12479 (39.0)	18511 (45.5)	18396 (52.4)	65632 (39.8)
Physical inactivity, N (%)	22961 (63.8)	13722 (65.5)	22112 (69.1)	30374 (74.7)	26889 (76.5)	116058 (70.5)
Stress (log transformed), mean (SD)	1.2 (0.7)	1.0 (0.8)	1.0 (0.7)	0.8 (0.7)	0.9 (0.7)	1.0 (0.7)
Health score, mean (SD)	80.8 (18.1)	84.3 (15.8)	86.4 (14.8)	89.1 (12.7)	89.6 (12.0)	86.3 (15.1)

Table 4.3 Marginal effect estimate (95% Confidence Interval)a for educational attainment on health using g-computation formula[b], World Health Survey 2002-2004 (N=164,743).

	No formal education	**Less than primary school**	**Primary school completed**	**Secondary school completed**	**High school or beyond**
Total effect	-4.05 (-4.24, -3.85)	-3.15 (-3.35, -2.95)	-2.09 (-2.24, -1.94)	-1.06 (-1.19, -0.92)	Reference
Natural direct effect	-3.08 (-3.31, -2.85)	-2.28 (-2.49, -2.06)	-1.50 (-1.65, -1.34)	-0.74 (-0.87, -0.61)	Reference
Natural indirect effect-A	-0.54 (-0.68, -0.41)	-0.49 (-0.60, -0.37)	-0.32 (-0.38, -0.26)	-0.18 (-0.21, -0.15)	Reference
Natural indirect effect-B	-0.43 (-0.49, -0.37)	-0.39 (-0.45, -0.34)	-0.28 (-0.31, -0.24)	-0.14 (-0.16, -0.12)	Reference
Controlled direct effect-00	-0.10 (-0.40, 0.19)	-0.56 (-0.88, -0.24)	-0.22 (-0.46, 0.02)	-0.34 (-0.57, -0.12)	Reference
Controlled direct effect-X0	-1.12 (-1.38, -0.86)	-1.20 (-1.46, -0.94)	-0.70 (-0.91, -0.50)	-0.55 (-0.75, -0.35)	Reference

[a] Wald type confidence intervals (CIs) were calculated as: point estimate ± 1.96 × SD, where SD was the standard deviation of the 200 point estimates from 200 bootstrapped samples.

[b] g-computation formula approach was domestic product per capita. implemented via Monte Carlo simulation, accounting for confounding due to age, gender, WHO region and country level wealth, measured by gross

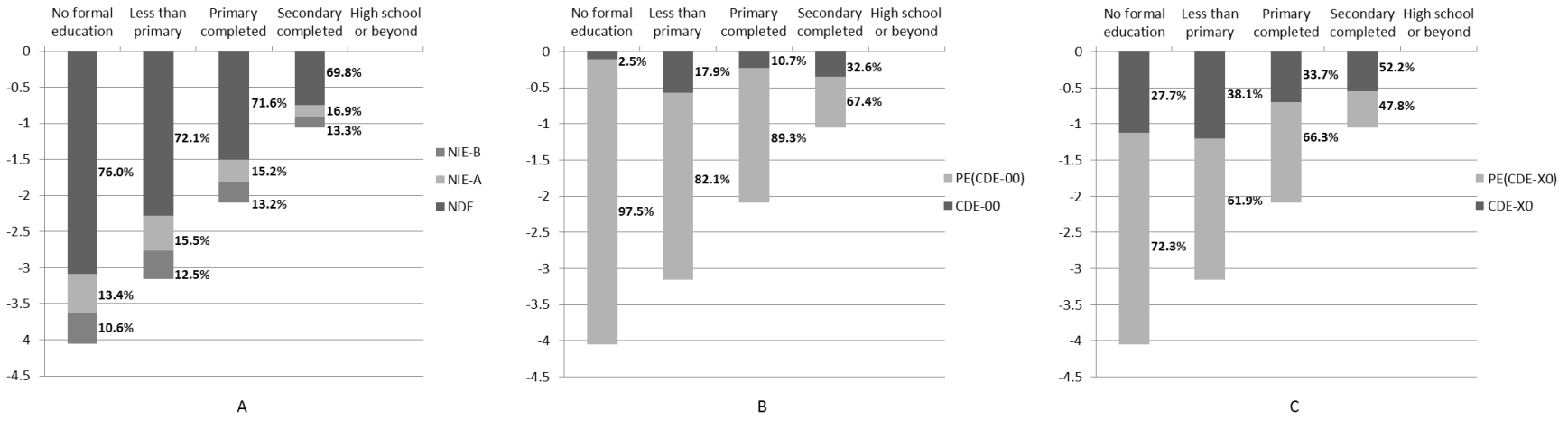

Figure 4.3 Effect decomposition and proportion explained by each pathway. A: natural decomposition; B and C: decomposition involving controlled direct effect. NDE: natural direct effect; NIE-A: natural indirect effect that is through social mediating factors and their consequences; NIE-B: natural indirect effect that is through health behavioral mediators only; CDE-00: controlled direct effect while fixing all mediators at the reference level (i.e. living in urban areas, being employed and married, not currently smoking, never used alcohol, being physically active, and not feeling stressed); CDE-X0: controlled direct effect while fixing only health behavioral mediators at the reference level (i.e. not currently smoking, never used alcohol, being physically active, not feeling stressed); PE(CDE-00) and PE(CDE-X0): the corresponding complement of total effect, also called "proportion eliminated".

4.5 Discussion

This large population-based global study found a widening gap in health status among participants with varying educational attainment compared to participants with high school or beyond education. Mechanistically, the major contribution of the negative impact of lower education on health was through pathways other than through social factors or health behaviors. Still, pathways through social factors or health behaviors accounted for more than one tenth of the health disparities due to education. A substantial amount of the observed health disparities would be eliminated if, in addition to increasing education, everyone had healthy behaviors and achieved the desired level of social factors via hypothetical interventions.

We found that health behaviors contributed to the education 'gradient' in health, though the former did not fully explain the latter. This is in line with one study that found a significant direct effect of education even adjusting for work and economic conditions, social-psychological resources, and health lifestyle[99], but not with the other, which found no educational impact after adjusting for income and health behaviors.[107] From a mechanistic perspective, only 11% to 13% of the educational disparity was attributed to the pathway from education to health behaviors and in turn to health. Direct comparison to the existing literature is difficult due to the different methods used in defining and estimating the pathway effects. Only one other study used causal mediation analysis to

examine the mediating role of health behaviors in the relation between education and diabetes incidence.[117] Body mass index and physical activity appeared to be mediating such relationship but the mediation proportion cannot be calculated because some of the pathways operated in the opposite directions. The large portion of direct effect not explained in the natural decomposition could be attributed to other important pathways such as physical and social environment, access to health care, psychosocial factors such as job control or social support.[90,109,118]

Health behaviors and social factors may be less important mechanistic mediators in the current study sample, but they were important effect modifier, as can be seen in the discrepancy between natural and controlled direct effects (DEs). In our study, interaction was present between education and each mediator in affecting health. In this case, the three types of DEs examined in the current study can differ because they captured the direct impact of education on health under different assignments of the mediators. The NDE evaluated the education effect had the distribution of the social factors and health behaviors for all participants achieved the same distributions as those among participants with high school or beyond education. In other words, it quantified the remaining health disparity due to education had all participants achieved the same living status and behave the same way as people with high school or beyond education did in terms of residence, employment, marriage, smoking, alcohol drinking, physical activity, and stress level. On

the other hand, the CDE-00 evaluated the remaining health disparity due to education had everyone lived in urban areas, been employed and married, never smoked nor used alcohol, and been physically active and not stressed. Similarly, the CDE-X0 quantified the remaining education effect had everyone never smoked nor used alcohol, and been physically active and not stressed. Multiple values of CDE-00 and CDE-X0 were possible, depending on the value we set the mediators. The presence of such education-mediator interaction highlights the need for health behavior interventions in addition to the effort in increasing education, in that these behavioral factors not only mediate the education impact but also moderate the strength of such impact. Health gap by education will exacerbate in the presence of unhealthy behaviors (appendix Table A 4.3).

Another reason for such difference was that these social factors and health behaviors were not completely deterministic by educational attainment: the mediator assignments were quite different for natural versus controlled direct effects. Consider unemployment as an example: despite the lower rates seen among the most educated, the unemployment rate was far from zero (data not shown), an ideal scenario that was evaluated in the CDE-00. Also, the most educated were not the ones that had the healthiest profiles; they drank alcohol more and were more physically inactive. Some of the positive impact of reducing smoking rates or stress may be offset by the negative impact of more alcohol use and being physically inactive and thus the NDE differed from the two CDEs.

From a public health intervention perspective, 48% to 72% of the educational gradient in health can be prevented by setting health behaviors at the desired level. Using the difference method, the British Whitehall II studies reported that health behaviors assessed at baseline explained 42%, 29%, and 61% of the socioeconomic gradient (measured by occupational grade) in all-cause, CVD, and non-cancer/non-CVD mortality whereas the repeated assessments of these behaviors during follow-up explained 72%, 45%, and 94% respectively.[114] In a later analysis of the Whitehall II study in comparison to the French GAZEL study, health behaviors were found to attenuate the association of SES with mortality by 75% in the former but only by 19% in the later.[106] Using data from the National Health and Nutrition Examination Survey, researchers found that in the low-income group, health behaviors attenuated the risk of all-cause and CVD/diabetes mortality by 30% and 21%, respectively.[113] In a study analyzing data from the National Health Interview Survey, the effect of education on mortality was reduced by 30% when controlling for exercise, smoking, drinking, seat belt use, and use of preventive care.[110] Our study is not easily comparable to the above studies because of the different behavioral factors included, health outcomes, and measures of SES. Some scholars did point out that different measures of SES are not interchangeable[119,120] and there is variation in the SES-health association because of the choice of measure.[121] It is also possible that the causal chain from SES to health/mortality via health behaviors played

out differently due to the difference in social patterning of unhealthy behaviors between countries.[106] Nevertheless, our study showed that health behaviors played an important role, especially among the least educated. Under hypothetical intervention of fixing the health behaviors at the desired level, the health disparity gap by education narrowed as educational attainment increased. People with no formal education would potentially benefit most from interventions that promote healthy behaviors in terms of narrowing the educational gradient in health.

To the best of our knowledge, this is the first study to quantify the contribution of underlying pathways that explained educational disparities in health across countries and continents using causal inference technique. The use of standardized global health data allowed for pooling data from multiple high-, middle-, and low-income countries and examining a global picture. We used causal mediation analysis tool that incorporated nonlinear relationships, which is crucial in the presence of exposure-mediator interaction. We presented results from both mechanistic and interventional perspectives that shed light on the well-established yet mysterious relationship between education and health. The hierarchical nature of the data was accounted for by using multilevel generalized linear models. Apart from partitioning the impact of education on health into pathway-specific effects, we also examined the remaining health disparities due to education under

hypothetical intervention of either setting health behaviors singly or combined with social factors at the desired levels.

Several methodological limitations need to be addressed. Due to the cross-sectional nature of the WHS data set, we assumed that educational attainment preceded participants residence, employment and marital status, health behaviors including current smoking status, alcohol use, physical activity status and stress level, and the present health status. Also, social factors were assumed to precede health behaviors, which preceded current health status. This is a reasonable assumption after we restricted our sample to participants aged 25 and older. Education, unlike income,[122] is less likely to be influenced by mid-life health conditions. Sensitivity analysis that restricted analysis to participants aged 40 and older (N=91,728) revealed similar patterns but slightly larger estimates. Repeated measurements on behaviors were not available, which explained a significantly greater part of the SES-mortality association compared to baseline-only assessments.[114] Due to missing information on health behaviors and health status, we lost participants from 24 countries, most of which were countries from the European Region. We also did not include fruit and vegetable consumption in our analysis due to vast missing values. We did not impose directionality between different social factors nor between different health behaviors; rather, we hypothesized variables within each of these two constructs were related by their upstream determinants as depicted in our DAG. Despite the use of

causal inference techniques, our result could still be subject to uncontrolled confounding between health behaviors and health status and measurement error biases. The results of CDEs should be interpreted with caution. They correspond to an ideal scenario that might never happen: you cannot force people to be married or have no stress. Therefore, they can be an overestimation of the educational disparities in health that could be eliminated by such joint interventions on health behaviors and social factors. In future studies, we will explore different intervention scenarios and the combinations of them in reducing health disparities by education.

This study provided evidence on the contribution of underlying pathways that explained educational gradient in health. Mechanistically, the impact of education on health was mainly direct. Yet, if the population can achieve the desired levels for health behaviors by certain interventions, a large portion of educational disparities could be eliminated, especially among those with no formal education. The results showed that natural direct effect and controlled direct can differ substantially in the presence of exposure-mediator interactions, which should be taken into account in future studies. Also, our study highlighted the need for continuing efforts on health behavior interventions among the less educated as countries throughout the world continue to achieve universal primary education or universal secondary education in the post-2015 era of the Millennium Development Goal.[123]

4.6 Appendix

Description of the World Health Survey (WHS) data set

Within each country, samples were probabilistically selected with every individual being assigned to a known non-zero selection probability. These samples were nationally representative except in China, Comoros, Congo, Côte d'Ivoire, India, and the Russian Federation, where the survey was carried out in geographically limited regions. This study included participants from 17 countries in the African region, 10 in the European region, six in the Americas, five in the South-East Asia region, five in the western Pacific region, and three in the Eastern Mediterranean region (Table A 4.1). All respondents were interviewed face-to-face with the standardized WHS survey, which included questions regarding demographic, socioeconomic, and behavioral factors.

Assumptions for identification

To estimate the effects defined above using the observational data, we assumed stable unit treatment value assumption (SUTVA),[46,93] general consistency, conditional exchangeability (no uncontrolled confounding), and positivity.[53] The conditional exchangeability assumption for natural effects included: (i) no uncontrolled confounding of the (X, $\mathbf{M_A}$, $\mathbf{M_B}$) – Y, X – $\mathbf{M_A}$, or (X, $\mathbf{M_A}$) – $\mathbf{M_B}$ relations given covariate set $\mathbf{Z}$, and (ii) no members of the covariate set $\mathbf{Z}$ are affected by X *or* $\mathbf{M_A}$. To identify controlled direct

effect, we assumed no uncontrolled confounding of the (X, $\mathbf{M_A}$, $\mathbf{M_B}$) – Y relationship given $\mathbf{Z}$ (CDE-00) and of the (X, $\mathbf{M_B}$) – Y relationship given $\mathbf{Z}$ (CDE-X0) respectively.

Description of the g-computation steps

We implemented the parametric g-formula algorithm in the three steps described here. First, we used multilevel generalized linear models with random intercept for country for each of the social factors and individual health behaviors and health score to account for the clustering within country (PROC MIXED procedure for stress and health score and PROC GLIMMIX procedure for the other variables in SAS). For each prediction model, confounders including an age-squared term and preceding factors for the corresponding outcome are included based on Figure 1. Bivariate interaction terms between education and all individual level factors were included as covariates in prediction models if the terms were significant at the $P<0.15$ level. Second, we created five copies of the original sample and assigned each copy the following education level: “no formal education”, “less than primary school”, “primary school completed”, “secondary school completed”, or “high school and beyond”. The pooled data set contained five synthetic cohorts under different education interventions ($X^{INT}=x^{INT}$). We then simulated potential variables for (i) social factors, (ii) health behaviors, and (iii) health sequentially using both the fixed effect estimates and the random intercept value for each country obtained from the first

step. Use no formal education as index intervention as an example. According to the counterfactual definitions presented in Table 4.1, we simulated:

(i) each potential social factor had we assign everyone no formal education [$\mathbf{M_A}(x)$], or high school or beyond education [$\mathbf{M_A}(x^*)$], or under the specific education intervention for their own cohort [$\mathbf{M_A}(x^{INT})$, i.e. combinations of $\mathbf{M_A}(x)$ and $\mathbf{M_A}(x^*)$];

(ii) potential health behavior variables under different education intervention and potential social factor assignments: $\mathbf{M_B}(x^{INT}, \mathbf{M_A}(x^{INT}))$, $\mathbf{M_B}(x^*, \mathbf{M_A}(x^*))$, $\mathbf{M_B}(x^*, \mathbf{M_A}(x^{INT}))$, and $\mathbf{M_B}(x^{INT}, \mathbf{M_A}(x))$; and

(iii) potential health status under education intervention, potential social factor from (i) and potential health behaviors from (ii): $Y(x^{INT}, \mathbf{M_A}(x^{INT}), \mathbf{M_B}(x^{INT}, \mathbf{M_A}(x^{INT})))$ (TE), $Y(x^{INT}, \mathbf{M_A}(x^*), \mathbf{M_B}(x^*, \mathbf{M_A}(x^*)))$ (NDE), $Y(x, \mathbf{M_A}(x^{INT}), \mathbf{M_B}(x^*, \mathbf{M_A}(x^{INT})))$ (NIE-A), $Y(x, \mathbf{M_A}(x), \mathbf{M_B}(x^{INT}, \mathbf{M_A}(x)))$ (NIE-B), $Y(x^{INT}, \mathbf{M_A}=\mathbf{m_a}, \mathbf{M_B}=\mathbf{m_b})$ (CDE-00), $Y(x^{INT}, \mathbf{M_A}(x^{INT}), \mathbf{M_B}=\mathbf{m_b})$.

To reduce Monte Carlo simulation error, the simulation was done on a dataset 200 times the size of the original (obtained via resampling with replacement), but the parameter estimation were based on the original sample size. For continuous stress and health score, the simulated value was bounded within the observed values ($0 \leq \text{stress} \leq 2.32$ and $0 \leq \text{health} \leq 99.9$). The final step involved regressing each potential health variable from

(iii) on education intervention to obtain the point estimate for the corresponding marginal effect.

We repeated the above three steps on 200 bootstrapped samples taken at random with replacement from the original data by country. The Wald type 95% confidence interval (CI) was calculated as: point estimate ± 1.96 × SD, where SD was the standard deviation of the 200 resultant point estimates from the final regression in the third step.

Appendix Tables

Table A 4.1 Country-specific sample size, percent female, mean age, and national 2003 Gross Domestic Product per capita (GDP/c), World Health Surveys 2002-2004.

Country	Initial sample size	N missing health score	N missing education information	N missing demographics	N missing health behaviors	Final sample size	Female (%)	Mean age	GDP/c (current USD)
African Region (AFR)									
Burkina Faso	3607	91	0	4	25	3486	50.7	41.4	332
Chad	3628	364	0	43	281	2991	51.8	41.8	292
Comoros	1411	55	0	82	21	1262	57.0	47.5	569
Congo	1937	414	11	247	540	1212	52.4	40.0	1039
Côte d'Ivoire	2402	248	13	86	117	2027	42.0	40.3	905
Ethiopia	3775	442	2	5	254	3115	51.1	41.9	117
Ghana	3302	114	34	25	96	3047	55.7	45.1	603
Kenya	3449	47	0	5	65	3332	57.9	42.6	504
Malawi	3761	126	1	16	39	3519	56.9	42.3	262
Mali	3176	2450	15	541	646	514	43.7	46.2	376
Mauritania	3011	295	5	114	604	2108	61.7	43.2	527
Mauritius	3385	302	0	2	24	3066	52.7	45.2	4830
Namibia	3284	1201	3	205	255	2016	59.3	42.6	2489
Senegal	2542	553	9	609	769	1247	48.2	42.9	643
South Africa	1876	331	1	39	89	1444	53.1	41.8	3739
Swaziland	2396	870	8	786	897	1390	54.0	43.8	1704
Zambia	2847	333	0	2	15	2490	53.5	41.2	399
Zimbabwe	3020	117	0	45	3020	0	64.8	43.1	529
Region of the Americas (AMR)									
Brazil	4209	548	0	139	4209	0	56.8	45.6	3039

Table A 4.1 Country-specific sample size, percent female, mean age, and national 2003 Gross Domestic Product per capita (GDP/c), World Health Surveys 2002-2004.

Country	Initial sample size	N missing health score	N missing education information	N missing demographics	N missing health behaviors	Final sample size	Female (%)	Mean age	GDP/c (current USD)
Dominican Republic	3758	61	1	5	67	3638	53.2	45.8	2210
Ecuador	3869	326	7	97	1855	1801	55.7	45.0	2442
Guatemala	3836	143	173	100	105	3351	61.1	44.6	1817
Mexico	32129	0	0	0	0	32129	57.5	45.1	6601
Paraguay	4062	45	0	1	30	3993	54.6	44.9	1159
Uruguay	2680	22	0	3	17	2640	51.8	48.7	3622
Eastern Mediterranean Region (EMR)									
Morocco	4184	4184	0	257	233	0	58.3	44.9	1684
Pakistan	5030	192	1	29	546	4315	45.3	41.6	597
Tunisia	4213	344	0	27	495	3430	54.9	45.9	2788
United Arab Emirates	984	65	3	0	64	863	47.8	40.3	36906
European Region (EUR)									
Austria	940	940	0	17	940	0	62.5	48.0	32019
Belgium	875	875	0	107	875	0	56.2	48.8	30675
Bosnia and Herzegovina	917	386	0	1	5	526	58.3	50.1	2182
Croatia	932	20	0	1	11	902	59.9	54.1	7857
Czech Republic	828	90	1	5	11	729	55.6	51.3	9732
Denmark	959	959	1	0	959	0	53.0	52.2	40517
Estonia	928	41	0	3	10	876	63.7	52.3	7333
Finland	944	944	0	1	944	0	55.3	55.1	32814
France	889	889	0	34	889	0	60.6	46.6	29657
Georgia	2441	16	0	64	12	2352	57.9	52.2	874

Table A 4.1 Country-specific sample size, percent female, mean age, and national 2003 Gross Domestic Product per capita (GDP/c), World Health Surveys 2002-2004.

Country	Initial sample size	N missing health score	N missing education information	N missing demographics	N missing health behaviors	Final sample size	Female (%)	Mean age	GDP/c (current USD)
Germany	1147	1147	1	35	1147	0	59.9	53.2	29864
Greece	916	916	0	2	916	0	50.1	53.8	18269
Hungary	1262	315	1	4	1262	0	59.4	53.0	8355
Ireland	866	866	5	113	866	0	56.0	48.5	40759
Israel	1075	1075	6	12	1075	0	58.1	48.3	19407
Italy	907	907	0	7	907	0	58.2	51.1	27135
Kazakhstan	4111	105	0	2	3	4001	65.8	43.3	2091
Latvia	763	72	0	1	763	0	68.3	54.6	5632
Luxembourg	620	620	0	3	620	0	50.8	48.3	65088
Netherlands	825	825	0	825	825	0	70.3	51.3	35385
Norway	872	872	2	872	872	0	50.8	50.9	50165
Portugal	911	911	0	0	911	0	62.5	54.4	15802
Russian Federation	4070	278	21	10	180	3605	64.6	54.0	2970
Slovakia	1922	680	573	598	613	1170	63.9	43.8	6307
Slovenia	512	30	1	512	4	0	54.3	50.9	14914
Spain	5959	187	0	5	31	5741	59.1	54.9	21583
Sweden	908	908	0	21	908	0	58.7	53.9	37030
Turkey	9678	344	7649	9678	9678	0	56.3	45.3	4595
Ukraine	2517	205	0	234	74	2048	65.3	50.8	1088
United Kingdom	1069	1069	0	79	1069	0	62.9	53.8	32561
South-East Asia Region (SEAR)									
Bangladesh	4528	821	1	9	49	3666	52.2	42.6	427
India	8140	1640	56	76	369	6238	51.7	43.0	541
Myanmar	4996	4	0	0	3	4989	57.3	44.6	200

Table A 4.1 Country-specific sample size, percent female, mean age, and national 2003 Gross Domestic Product per capita (GDP/c), World Health Surveys 2002-2004.

Country	Initial sample size	N missing health score	N missing education information	N missing demographics	N missing health behaviors	Final sample size	Female (%)	Mean age	GDP/c (current USD)
Nepal	6979	49	0	2	29	6900	56.3	43.3	264
Sri Lanka	5642	710	0	375	369	4303	54.0	44.9	968
Western Pacific Region (WPR)									
Australia	3316	3316	75	3316	3316	0	58.0	49.7	28017
China	3674	54	0	5	3	3614	51.4	47.2	1267
Lao People's Democratic Republic	4060	86	9	3	50	3919	52.7	41.8	358
Malaysia	5250	203	4	20	66	4996	56.8	44.2	4607
Philippines	8380	110	0	9	58	8207	54.6	42.6	1016
Viet Nam	2983	1428	0	8	26	1535	55.5	43.4	475

Table A 4.2 Characteristics of 66,531 participants excluded from the main analyses due to missing values in one or more of the variables, World Health Survey 2002-2004.

Characteristics	Sample size	Descriptive statistics
Total, N (%)		
Age, mean (SD)	66237	46.7 (15.6)
Females, N (%)	66474	37961 (57.1)
Educational attainment	57835	
No formal education		14595 (25.2)
Less than primary school		5593 (9.7)
Primary school completed		10882 (18.8)
Secondary school completed		11044 (19.1)
High school and beyond		15721 (27.2)
Social factors		
Living in rural areas, N (%)	59677	27165 (45.5)
Unemployment, N (%)	61638	33309 (54)
Not Married, N (%)	55639	20798 (37.4)
Individual health behaviors		
Currently smoking, N (%)	45080	10857 (24.1)
Alcohol use, N (%)	43458	14159 (32.6)
Physical inactivity, N (%)	40637	31416 (77.3)
Stress (log transformed), mean (SD)	42956	0.9 (0.8)

Note: Table A 4.3 below shows the parameter estimates from a regular linear mixed model for health score, conditional on education, compositional factors, health behaviors and stress, and covariates while allowing for bivariate product terms between education level and every mediator and individual level confounders. The result presented below cannot be directly compared to the marginal CDE-00 presented in the main text because those effects are marginalized over covariates, representing the population average. Due to the presence of interaction between education and sex, conditional estimates were presented for males and females separately.

Table A 4.3 Conditional controlled direct effect estimate (95% confidence interval) for educational attainment on health using linear mixed model with random intercept for countrya, World Health Survey 2002-2004.

	No formal education	**Less than primary school**	**Primary school completed**	**Secondary school completed**	**High school or beyond**
Males					
Conditional CDE-00[b]	0.24 (-0.38, 0.86)	-0.76 (-1.46, -0.07)	-0.47 (-1.07, 0.13)	-0.92 (-1.47, -0.36)	Reference
Conditional CDE-01[c]	-4.22 (-4.80, -3.63)	-2.86 (-3.53, -2.20)	-3.38 (-3.94, -2.83)	-3.34 (-3.84, -2.84)	Reference
Females					
Conditional CDE-00	-0.03 (-0.67, 0.61)	-1.28 (-2.00, -0.55)	-0.29 (-0.92, 0.34)	-0.49 (-1.09, 0.10)	Reference
Conditional CDE-01	-4.49 (-5.18, -3.80)	-3.38 (-4.19, -2.57)	-3.21 (-3.89, -2.52)	-2.92 (-3.54, -2.30)	Reference

[a] Model included bivariate product terms between education level and every mediator and individual level confounder.
[b] Conditional CDE-00 represents the controlled direct effect when fixing both social factors and health behaviors at reference levels (i.e. living in urban areas, being employed and married, not smoking, never used alcohol, being physically active and not stressed) for 45 year-old participants.
[c] Conditional CDE-01 represents the controlled direct effect when fixing social factors at reference levels (i.e. living in urban areas and being employed and married) but health behaviors at index levels (i.e., smoking, ever used alcohol, being physically inactive and having 1-unit increase in stress score) for 45 year-old participants.

Chapter 5. The Modifying Influence of Leisure Time Physical Activity on the Impact of Occupational Physical on 20-Year Incidence of Acute Myocardial Infarction

5.1 Abstract

Objectives: To disentangle the complex interplay between occupational physical activity (OPA) and leisure-time physical activity (LTPA) in affecting cardiovascular health, this study aimed to examine (1) interactions between OPA and LTPA and their combined effect on 20-year incidence of acute myocardial infarction (AMI), and (2) the effect of OPA on AMI that is mediated through LTPA.

Methods: We analyzed data on 1891 men, aged 42-60 years, from the prospective Kuopio Ischemic Heart Disease Risk Factor Study. We obtained first-time incident AMI after baseline via hospitalization discharge and death registries. OPA was measured as relative aerobic strain (RAS), which took into account both the absolute energy expenditure at work and the workers' cardiorespiratory fitness. Averaged 12-month LTPA was assessed based on survey as were with potential confounders.

Results: We found multiplicative interactions between OPA and LTPA among men with IHD. The multivariable-adjusted (age, education, smoking, alcohol consumption, technical and psychosocial job factors) Cox model showed that high OPA positively predicted AMI at low LTPA level for both men without and with IHD —HR 1.27 (95% CI: 0.96, 1.68) and HR 1.59

(95% CI: 0.99, 1.68) respectively—but not at high levels of LTPA. Analysis using continuous physical activity measures showed a stronger effect estimate in men without IHD. The combination of high OPA and low LTPA constituted the group associated with the highest risk for AMI, irrespective of IHD status. LTPA was not independently predictive of AMI and did not mediate the impact of OPA on AMI.

Conclusions: LTPA interacted with OPA on the multiplicative scale but did not mediate the effect of OPA on AMI.

5.2 Introduction

Sedentary lifestyle, or physical inactivity, is an established risk factor for cardiovascular disease.[124–126] Accordingly, physical activity both in work setting and leisure time have been recommended.[127] While leisure-time physical activity (LTPA) has been well documented to promote health,[128,129] the effect of occupational physical activity (OPA) is inconsistent.[130] Without adjustment for LTPA, higher levels of OPA were reported to be protective against CVD in some studies,[131,132] have no effect,[133,134] or increase the CVD risk.[135,136] When adjusting for LTPA, some studies showed that greater OPA is associated with progression of carotid atherosclerosis,[137] and increased AMI incidence[138] or risk of IHD mortality.[139]

One explanation for such inconsistency could be that the effects of OPA depend on the level of LTPA[140,141] and possibly individual aerobic fitness.[137,142] If interaction between OPA and LTPA

exists, a model ignoring such interaction could in some scenarios result in cancellation of effect of one variable across levels of another and yield a misleading average estimate of no effect. Another explanation for such pattern could be that high levels of OPA preclude workers from engaging in LTPA; thus these workers cannot benefit from LTPA. Finally, OPA could directly affect CVD. Based on the negative correlation between OPA and LTPA observed in previous work[137] and previous observations that LTPA participation was relatively low among blue-collar workers,[143] we hypothesized that LTPA both interacts with OPA and mediates the effect of OPA on CVD. Therefore, as a follow-up study to our previous publications examining the relationship between OPA or LTPA and cardiovascular outcomes, we further assess the modifying and mediating role of LTPA on the pathway from OPA to 20-year incidence of AMI. We also explored the impact of LTPA on AMI at different levels of OPA. We conducted separate analyses for men with and without preexisting ischemic heart disease (IHD), as past studies suggested a heterogeneous OPA effect by IHD status.[138]

More specifically, and separately for men with and without preexisting IHD, our study aimed to: (1) assess both multiplicative and additive interaction between OPA and LTPA, and their combined effect on 20-year incidence of AMI, and (2) examine the potential mediating role of LTPA on the pathway from OPA to AMI, using causal mediation analysis that allowed for exposure-mediator interaction.[28,144]

5.3 Methods

Study design, setting and population

Participants were from the prospective Kuopio Ischemic Heart Disease Risk Factor (KIHD) Study, an age-stratified, random, population-based sample of Eastern Finnish men, residing in the city of Kuopio or its surrounding rural communities. Details of the study population are available elsewhere.[138,145] Out of 3235 eligible men aged 42, 48, 54, or 60 years, 2682 (82.9%) men agreed to participate, with 553 men being excluded due to refusal (N=367) and no contact (N=186). All participants underwent baseline examinations and interviews between March 1984 and December 1989 and were passively followed by national hospitalization discharge and death registries until 2011. We excluded 791 participants who were not working at baseline or in the 12 months prior, resulting in a final study sample of 1891 participants with complete information on all the baseline covariates for the main analyses. All participants provided written informed consent. The University of California, Los Angeles (UCLA) Institutional Review Board approved this study.

Assessment of incidence of acute myocardial infarction

As described previously,[138] we ascertained first-time incident AMI (ICD-9 code 410) during follow-up via record linkage with national hospitalization discharge and death registries including the national AMI register established under the World Health Organizations "Monitoring of Trends and Determinants of Cardiovascular Diseases (MONICA)" project.[146,147]

A university-based cardiologist for this study confirmed hospital discharge diagnoses using other hospital records, lab results, and electrocardiograms. We censored the follow-up at 31 December, 2011 or date of death whichever came first.

Assessment of occupational physical activity

We measured OPA as relative aerobic strain (RAS), the most predictive factor for AMI among other OPA measures found in the same study population.[148] RAS (aka relative aerobic workload) expresses the caloric demands of work as a percentage of the individual worker's aerobic cardiorespiratory fitness or maximal work capacity (%VO_2max).[149] RAS takes into account both the absolute energy expenditure (EE) and the workers' individual aerobic capacity. Detailed descriptions of the assessment of these variables can be found elsewhere.[148,149] Based on work physiology and ergonomic principles, it is often the misfit between high job-related energy demands and low worker aerobic capacity, rather than a high absolute amount of EE alone, that will lead to elevated blood pressure and heart rate during work, two established risk factors for AMI.[148] Also, OPA has been shown to be detrimental among workers with low cardiorespiratory fitness but not among those with high fitness level.[150] Thus, we decided to use RAS as our OPA exposure measure in the main analyses.

Absolute EE at work (in kcal/day) was assessed from baseline interview data on time spent in various activities at work during a typical workday and reference data on the energy

requirements (kcal/kg/hour) of these activities. EE in kcal for each reported activity was calculated by multiplying the duration (hours per day) by the respective intensity (MET) and body weight (kg) for each individual. EE per typical workday was the sum of EE for all activities.

Cardiorespiratory fitness (also known as aerobic capacity or VO_2max) was measured with a maximal, symptom-limited exercise-tolerance test on a bicycle ergometer as explained in detail elsewhere.[151–153] VO_2max, in ml O_2 per kg per minute, was defined as the highest value or plateau in oxygen uptake during maximal symptom-limited bicycle ergometer and was standardized by body weight.

Assessment of leisure-time physical activity

LTPA was measured using the KIHD 12-Month Leisure-Time Physical Activity History, a modified version of the Minnesota Leisure Time Physical Activity questionnaire,[154] that included the 16 most common leisure time physical activities of middle-aged Finnish men.[153,155] Respondents were asked to record the frequency, duration, and intensity of each of 16 activities performed for each of the 12 previous months. Conditioning (of vigorous-intensity) LTPA included walking, jogging, cross-country skiing, bicycling, swimming, rowing, ball games, and gymnastics, dancing, or weightlifting. We calculated the sum of these activities and obtained the

average conditioning LTPA, expressed in minutes per week, in the previous year. Unless otherwise noted, we used LTPA to represent conditioning LTPA throughout the article.

Assessment of covariates

We included as confounders age, education, participation in unrelated lipid-lowering drug trial (placebo group, treatment group, versus none), and baseline IHD. Education was categorized into: (i) some elementary school, (ii) elementary school completed, or elementary school plus some junior high school, (iii) junior high school completed, or junior high school plus some senior high school, and (iv) senior high school completed or beyond. A continuous smoking variable "cigarette-years" was calculated based on the number of cigarettes per day and the number of years smoked. Alcohol consumption (grams per week) accounted for frequency of drinking and amount of drinks per occasion for each type of alcoholic beverage (beer, wine, spirits) for the last 12 months. *Psychosocial job factors* were measured using questionnaires that captured mental strain at work (11 items of psychological demands), social support at work (3 items), and stress from work deadlines. These factors have been associated with progression of atherosclerosis and an increased risk for myocardial infarction and mortality in this study population and showed satisfactory Cronbach's α coefficients.[156,157] We classified participants as having preexisting IHD at baseline if they (i) had a history of prior (before baseline) myocardial infarction or angina pectoris, (ii) currently used anti-angina medication, or (iii) had positive findings of angina according to the London School of Hygiene cardiovascular questionnaire.[158]

Statistical analysis

We summarized the participants' characteristics by their baseline IHD status.

Interaction analysis

We used Cox proportional hazard models[159] with adjustment of covariates listed in Table 5.1. We added an OPA × LTPA product term to assess the interaction between OPA and LTPA on the multiplicative scale. We also calculated the relative excess risk for interaction (RERI) as a measure for additive interaction:[160] $HR_{11} - HR_{10} - HR_{10} + 1$, where HR_{11}, HR_{10}, and HR_{01} respectively represented the joint effect of OPA and LTPA, the main effect of OPA, and the main effect of LTPA. Variables were recoded jointly when necessary so that the reference combined category represented lowest risk group.[161] OPA was modeled both as a binary indicator (RAS > 33% as high versus RAS ≤ 33% as low), based on the maximum level of 33% VO_2max recommended for 8 hours of work,[149,162] and continuous variable (1 unit representing a 20% increase in RAS) centered at a level of 23.5%. Similarly, LTPA was modeled as a binary indicator (LTPA≥ 75 minutes/week as high versus LTPA < 75 minutes/week as low) based on WHO global recommendations[163] as well as continuous variable (1 unit representing a 75 minutes/week increase). Binary OPA was paired with binary LTPA whereas continuous OPA was paired with continuous LTPA in all analyses. We also performed the analyses using low OPA and high LTPA as the reference category for binary physical activity measures and

compared different combination of OPA and LTPA relative to this reference group as done in the existing literature. We reported hazard ratio (HR) for AMI associated with a 1-unit increase in OPA and its corresponding 95% confidence interval (CI) at different levels of LTPA. Similarly, effect estimates associated with a 1-unit increase in LTPA at different levels of OPA were also presented. The quadratic term for the continuous RAS measure was not significant at P=0.1 level and did not improve model fit. Thus, the hazard function was modeled in a linear form for this measure.

Mediation analysis

We assumed that baseline OPA level determined the baseline LTPA level, not the other way around, and that first-time incidence of AMI occurred during follow-up can be attributed to the OPA and LTPA levels measured at baseline. We invoked the stable unit treatment value assumption (SUTVA),[46] and assumptions of consistency, positivity, conditional exchangeability (no-uncontrolled-confounding),[47,48] and no selection bias and measurement error. Further discussions of these assumptions can be found elsewhere.[164] We used recently proposed inverse-probability weighted (IPW) fitting of marginal structural models (MSMs) for causal mediation analysis[144] to estimate the marginal pure direct effect (PDE) of baseline OPA on AMI and the marginal total indirect effect (TIE) of baseline OPA via baseline LTPA (Figure 5.1). PDE was defined as the hazard ratio comparing high to low OPA levels while allowing LTPA attain the natural value under the low OPA level. TIE was defined as the hazard ratio comparing two

LTPA levels – the natural LTPA level under high OPA versus the natural LTPA level under low OPA – while setting OPA level to be high. Methodological details can be found in the appendix.

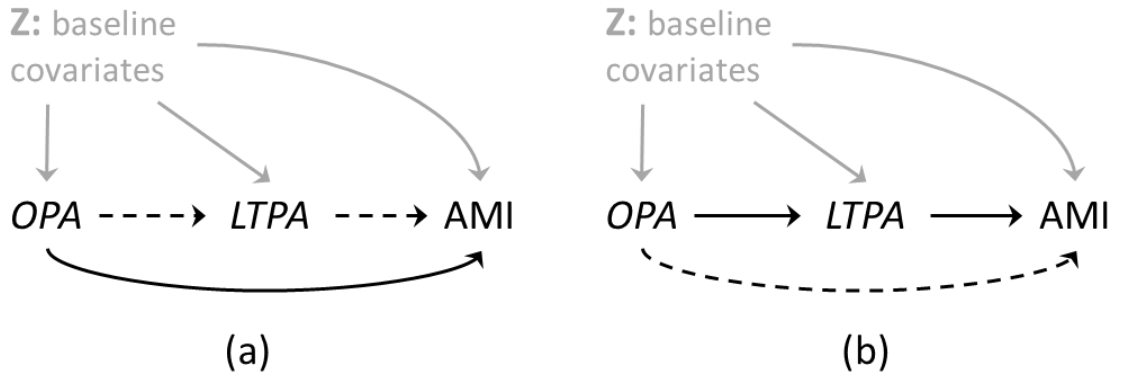

Figure 5.1 Graphical presentation (solid black lines) of pure direct effect (a) and total indirect effect (b) of occupational physical activity (OPA) on acute myocardial infarction (AMI), with mediator leisure-time physical activity (LTPA).

Sensitivity analysis

We conducted the following sensitivity analyses to test the robustness of our results for assessing interaction between OPA and LTPA. First, we repeated our analyses using trichotomized OPA (low: RAS≤23%, moderate: 23%≤RAS<33%, high: RAS>33%) and LTPA (low: LTPA<20 minutes/week, moderate: 20 minutes/week ≤LTPA<75 minutes/week, high: LTPA≥75 minutes/week) measures. Second, we repeated our main analyses with additional adjustment for biological factors including blood glucose, plasma fibrinogen, body mass index, LDL-cholesterol, HDL-cholesterol, systolic blood pressure, lipid-lowering mediation, and anti-hypertensive medication. Third, we used continuous absolute EE (500 kcal increase), centered at

the population mean of 2111 kcal/day, as an alternative measure of OPA to assess the interaction between OPA and LTPA.

For assessing the mediating role of LTPA, we additionally included 4-year LTPA as a second mediator and examined the effect of OPA via pathways involving baseline LTPA, or 4-year LTPA, or neither. We further restricted our analytical sample to 455 men without baseline IHD and who had complete information on all variables. Detailed sample restriction criteria, methodology, effect definition, and implementation steps can be found elsewhere[28] and in the appendix.

All analyses were performed using Stata version 14 (StataCorp LP, College Station, Texas).

5.4 Results

Characteristics of the study sample

The distribution of exposure variables and covariates by preexisting IHD status is listed in Table 5.1. Participants' mean age was 51.5 years [standard deviation (SD): 5.0] for participants without IHD and 53.5 (SD: 3.9) for those with IHD. Over 70% of the participants completed elementary school but not junior high school. Participants with IHD had higher levels of RAS, absolute EE, and mental strain at work, smoked and drank more, and experienced more stress from work deadlines, but had lower level of fitness than men without IHD. LTPA and social support at work were similar in these two subgroups.

Incidence of AMI

During an average of 19.56 years of follow-up (SD: 7.53; range: 0.01-27.76) and a total person-time of 36991 years, 495 first-time incident AMI occurred among 1891 study participants, yielding a yearly incidence rate of 1.34%. Among 1565 men without baseline IHD, 353 AMI occurred (yearly incidence 1.11%) whereas among 326 men with baseline IHD, 142 AMI occurred (yearly incidence 2.60%).

Interaction between binary OPA and binary LTPA in affecting AMI

Table 5.2 displays the associations between one type of PA and AMI at different levels of the other PA type by preexisting IHD status and the joint association of both OPA and LTPA with AMI, using low OPA and high LTPA as reference. For men without IHD, RAS greater than 33% was only positively associated with AMI incidence among men with low LTPA and with age adjustment (HR: 1.34, 95% CI: 1.01, 1.76) but such association attenuated after adjusting for other factors (HR: 1.27, 95% CI: 0.96, 1.68). For men with IHD, RAS positively predicted AMI at low LTPA (HR: 1.59, 95% CI: 0.99, 2.57) but not at high LTPA (HR: 1.04, 95% CI: 0.61, 1.79). For both IHD subgroups, high LTPA was weakly negatively associated with AMI at high but not at low OPA level. Compared to men with low OPA and high LTPA, men with high OPA and low LTPA had the highest risk for AMI, irrespective of IHD status (HR: 1.33, 95% CI: 0.99, 1.78 for men without IHD; HR: 1.36, 95% CI: 0.84, 2.18 for men without IHD). Multiplicative

interaction between OPA and LTPA was observed among men with IHD (ratio of HR: 0.65, *P*=0.240) but no additive interaction was observed for both subgroups.

Results from sensitivity analysis using trichotomized PA measures are presented in appendix Tables A 5.1 and A 5.2. Moderate and high OPA, compared to low OPA, was positively associated with AMI at moderate and high LTPA levels among men without IHD but only at high LTPA level among men with IHD. LTPA did not predict AMI across OPA levels. The combined effect of moderate OPA and moderate LTPA was greater than the product of their separate effects among men without IHD. The combined effect of high OPA and moderate LTPA was greater than the product of their separate effects among men with IHD. When using the low OPA and moderate LTPA as the general reference group for men without IHD, those with moderate OPA and moderate LTPA had the highest risk for AMI (HR: 2.35, 95% CI: 1.43, 3.85). For men with IHD, those with low OPA and moderate LTPA had the highest risk for AMI (HR: 3.92, 95% CI: 1.20, 12.78), when compared to moderate OPA and low LTPA reference group. Negative (i.e., the combined effect of two PA measures being smaller than the sum of their separate effects) but uncertain additive interactions were found (1) comparing moderate OPA and low LTPA to low OPA and moderate LTPA among men without IHD, and (2) comparing low OPA and high LTPA to moderate OPA and low LTPA among men with IHD.

Results from sensitivity analysis that additionally adjusted for biological factors are presented in appendix Table A 5.3. Neither RAS nor LTPA predict AMI. We did not observe multiplicative or additive interactions between RAS and LTPA.

Interaction between continuous OPA and continuous LTPA in affecting AMI

Table 5.3 depicts the associations between one type of PA and AMI at different levels of the other PA by preexisting IHD status. For men without IHD, a 20% increase in RAS (from a reference RAS level of 23.5%) positively predicted AMI incidence at both LTPA of 0 minute/week (HR: 1.45, 95% CI: 1.19-1.75) and at LTPA of 75 minutes/week (HR: 1.49, 95% CI: 1.28-1.75). Weaker associations between RAS and AMI were found among men with IHD (HR: 1.25, 95% CI: 0.96, 1.64 at LTPA of 0 minute/week; HR: 1.32, 95% CI: 1.07, 1.62 at LTPA of 75 minutes/week). We found no association between LTPA and AMI across levels of RAS and no multiplicative interaction between RAS and LTPA.

Table A 5.4 depicts results from using absolute EE as OPA measure. High absolute EE was associated with AMI only at LTPA level of 75 minutes/week (HR: 1.06, 95% CI: 1.00-1.13) but not at a lower LTPA level for men without IHD and was not associated with AMI across LTPA levels among men with IHD. LTPA did not predict AMI or interact with EE at the multiplicative scale.

The mediating role of LTPA on the pathway from OPA to AMI

Table 5.4 depicts the pure direct effect of OPA on AMI and the total indirect effect via baseline LTPA, estimated using IPW fitting of MSMs. We observed similar effect estimates from all three methods that differed only in the way confounding was handled. For men without IHD, the estimate for total effect of OPA on AMI was 1.31 (0.99, 1.79) when marginalizing over all covariates and LTPA. The majority of such positive impact of OPA on AMI was through pathways other than through LTPA (PDE: 1.27, 95% CI: 0.93, 1.74). The effect estimates for men with IHD were similar to that among men without IHD.

Table A 5.5 in the appendix depicts results from sensitivity analysis that used also LTPA measured at 4-year follow-up and further decomposed the total effect into baseline LTPA pathway-specific effect, 4-year LTPA pathway-specific effect, and natural direct effect that is through neither baseline nor 4-year LTPA. In this restricted sample of men without IHD, OPA was not associated with AMI overall (TE: 1.08, 95% CI: 0.58, 2.02). Weak positive direct effect (NDE: 1.22, 95% CI: 0.57, 2.46) and weak negative indirect effect via baseline LTPA ($NIE_{baseline}$: 0.93, 95% CI: 0.69, 1.21) were observed. Follow-up LTPA at 4 years did not mediate the effect of baseline OPA on AMI ($NIE_{4\text{-}year}$: 0.96, 95% CI: 0.79, 1.05).

Table 5.1 Characteristics of the study sample and distribution of exposure and covariates by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

	Men without IHD (N = 1565)	**Men with IHD (N = 326)**
Occupational physical activity (OPA)		
Relative aerobic strain (%), mean (SD)	29.7 (12.1)	38.5 (16.4)
Binary relative aerobic strain (>33%), N(%)	490 (31.3)	179 (54.9)
Absolute energy expenditure (kcal/day), mean (SD)	2078.0 (874.7)	2272.0 (969.6)
Leisure-time physical activity (LTPA)		
Conditioning LTPA (minutes/week), mean (SD)	104.6 (117.1)	108.9 (144.6)
Binary conditioning LTPA (≥75 minutes/week), N(%)	742 (47.4)	143 (43.9)
Cardiorespiratory fitness (VO_2max, O_2/kg/minute)	32.8 (7.0)	27.5 (6.9)
Covariates		
Age at baseline (years), mean (SD)	51.5 (5.1)	53.5 (3.9)
Age group, N(%)		
42 years old	292 (18.7)	17 (5.2)
48 years old	274 (17.5)	47 (14.4)
54 years old	916 (58.5)	233 (71.5)
60 years old	83 (5.3)	29 (8.9)
Participation in lipid-lowering drug trial, N(%)		
Placebo group	135 (8.6)	28 (8.6)
Treatment group	136 (8.7)	27 (8.3)
Education, N(%)		
Some elementary school	113 (7.2)	41 (12.6)
Elementary school completed/some junior high school	1144 (73.1)	251 (77.0)
Junior high school completed/ some senior high school	150 (9.6)	28 (8.6)
Senior high school completed or beyond	158 (10.1)	6 (1.8)
Behavioral factors		
Smoking (cigarettes/day × years)	141.8 (290.3)	210.2 (348.1)
Alcohol consumption (g/week)	71.5 (111.8)	88.3 (196.4)
Psychosocial job factors		
Mental strain at work index	11.5 (6.3)	13.4 (7.1)
Social support at work score	6.5 (2.5)	6.5 (2.4)
Stress from work deadlines	357 (22.8)	105 (32.2)

Table 5.2 Hazard ratios (HR) and 95% confidence intervals (95% CI) for the main effect and joint effect of occupational physical activity (OPA) and leisure-time physical activity (LTPA) on 20-year incidence of acute myocardial infarction (N=495) when both types of physical activity were modeled as binary variables, by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

	Men without IHD (N=1565)			**Men with IHD (N=326)**		
	N	**Age-adjusted HR (95% CI)**	**Fully-adjusted[a] HR (95% CI)**	**N**	**Age-adjusted HR (95% CI)**	**Fully-adjusted[a] HR (95% CI)**
Binary RAS (RAS>33% versus RAS≤33%)						
Low LTPA (< 75 minutes/week)	--	1.34 (1.01, 1.76)	1.27 (0.96, 1.68)	--	1.52 (0.95, 2.42)	1.59 (0.99, 2.57)
High LTPA (≥75 minutes/week)	--	1.20 (0.83, 1.74)	1.10 (0.76, 1.60)	--	0.97 (0.58, 1.63)	1.04 (0.61, 1.79)
Binary LTPA (LTPA≥75 minutes/week versus <75 minutes/week)						
Low OPA (RAS ≤ 33%)	--	0.85 (0.66, 1.11)	0.95 (0.73, 1.25)	--	1.18 (0.70, 1.99)	1.17 (0.69, 2.01)
High OPA (RAS > 33%)	--	0.77 (0.53, 1.12)	0.83 (0.57, 1.21)	--	0.76 (0.48, 1.19)	0.77 (0.48, 1.23)
P for Multiplicative Interaction[b]	--	0.649	0.543	--	0.205	0.240
Combination of RAS and LTPA, using low RAS and high LTPA as the reference						
Low OPA and high LTPA	571	Reference	Reference	80	Reference	Reference
High OPA and high LTPA	171	1.20 (0.83, 1.74)	1.10 (0.76, 1.60)	63	0.97 (0.58, 1.63)	1.04 (0.61, 1.79)
Low OPA and low LTPA	504	1.17 (0.90, 1.52)	1.05 (0.80, 1.37)	67	0.84 (0.50, 1.42)	0.85 (0.50, 1.46)
High OPA and low LTPA	319	1.57 (1.19, 2.07)	1.33 (0.99, 1.78)	116	1.28 (0.83, 1.99)	1.36 (0.84, 2.18)
RERI (95% CI)[c]	--	0.19 (-0.37, 0.75)	0.18 (-0.32, 0.69)	--	-0.55 (-1.49, 0.38)	-0.54 (-1.51, 0.43)

[a] Model adjusted for age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

[b] *P* value for the RAS × LTPA product term.

[c] RERIs were measures for additive interaction. For men without IHD, RERIs were calculated as $HR_{\text{high OPA, low LTPA}} - HR_{\text{high OPA, high LTPA}} - HR_{\text{low OPA, low LTPA}} + 1$. For men with IHD, RERIs were calculated as $HR_{\text{high OPA, high LTPA}} - HR_{\text{high OPA, low LTPA}} - HR_{\text{low OPA, high LTPA}} + 1$, using low OPA and low LTPA as the reference group.

Table 5.3 Hazard ratios (HR) and 95% confidence intervals (95% CI) for the main effect and joint effect of occupational physical activity (OPA) and leisure-time physical activity (LTPA) on 20-year incidence of acute myocardial infarction (N=495) when both types of physical activity were modeled as continuous variables, by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

	Men without IHD (N=1565)		**Men with IHD (N=326)**	
	Age-adjusted HR (95% CI)	**Fully-adjusted[a] HR (95% CI)**	**Age-adjusted HR (95% CI)**	**Fully-adjusted[a] HR (95% CI)**
Continuous RAS (20% increase)				
LTPA = 0 minute/week	1.45 (1.21, 1.74)	1.45 (1.19, 1.75)	1.20 (0.93, 1.55)	1.25 (0.96, 1.64)
LTPA = 75 minutes/week	1.50 (1.30, 1.74)	1.49 (1.28, 1.75)	1.26 (1.04, 1.52)	1.32 (1.07, 1.62)
Continuous LTPA (75 minutes/week increase)				
RAS = 23.5%	0.96 (0.88, 1.04)	0.98 (0.90, 1.07)	0.95 (0.82, 1.09)	0.95 (0.82, 1.10)
RAS = 43.5%	0.99 (0.91, 1.08)	1.01 (0.93, 1.10)	0.99 (0.88, 1.11)	1.00 (0.89, 1.12)
P for multiplicative interaction[b]	0.362	0.458	0.571	0.494

[a] Model adjusted for age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

[b] *P* value for the RAS × LTPA product term.

Table 5.4 Hazard ratios (HR) and 95% confidence intervals (95% CI) for the pure direct effect (PDE) and total indirect effect (TIE) of occupational physical activity as measured by binary relative aerobic strain (with leisure-time physical activity as mediator) on 20-year incidence of acute myocardial infarction (N=495) for all men, stratified by preexisting ischemic heart disease (IHD) status, and estimated using inverse-probability weighted (IPW) fitting of marginal structural models (MSMs)[a], in the Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

	Men without IHD (N=1565)	**Men with IHD (N=326)**
Method	**HR (95% CI)[b]**	**HR (95% CI)[b]**
MSM[c]		
PDE	1.27 (0.93, 1.74)	1.28 (0.88, 1.90)
TIE	1.04 (0.88, 1.22)	1.04 (0.88, 1.27)
Total effect	1.31 (0.99, 1.79)	1.33 (0.97, 1.85)
Conditional MSM[d]		
PDE	1.20 (0.96, 1.55)	1.28 (0.87, 1.93)
TIE	1.02 (0.90, 1.15)	1.05 (0.90, 1.24)
Total effect	1.22 (1.02, 1.54)	1.35 (0.96, 1.91)
Doubly robust MSM[e]		
PDE	1.23 (0.93, 1.67)	1.30 (0.88, 1.95)
TIE	1.04 (0.90, 1.21)	1.04 (0.89, 1.24)
Total effect	1.28 (0.98, 1.70)	1.34 (0.97, 1.92)
Conditional total effect[f]	1.22 (0.97, 1.53)	1.34 (0.94, 1.91)

[a] Occupational physical activity was measured by binary relative aerobic strain (RAS) indicator (RAS>33% versus RAS≤33%) and leisure-time physical activity was dichotomized (≥75 minutes/week versus <75 minutes/week). Covariates included age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

[b] Bias-corrected and accelerated 95% confidence intervals (CIs) were obtained using 1000 bootstrap samples.

[c] IPW was created based on a weight for OPA (dealing with confounding) and a weight for LTPA (decomposing effect).

[d] IPW was created based on a weight for LTPA (decomposing effect) only. Conditional MSM included covariates to control for confounding.

[e] IPW was created based on a weight for OPA (dealing with confounding) and a weight for LTPA (decomposing effect). In the final MSM, covariates were adjusted for.

[f] Cox proportional hazard model included OPA, age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

5.5 Discussion

This 20-year follow-up study examined the interaction between OPA and LTPA in affecting AMI incidence and the pathway effect of OPA on AMI via LTPA among men with and without preexisting IHD. We found high levels of OPA positively predicted AMI at low LTPA level for both IHD subgroups; however, LTPA was not predictive of AMI after accounting for OPA. We found multiplicative interactions between OPA and LTPA among men with IHD but no additive interactions. LTPA did not appear to mediate the effect of OPA on AMI.

Our finding on the positive link between OPA and AMI, regardless of categorical or continuous PA measures being used, is in line with a previous study on OPA and AMI from the same cohort, despite slightly different covariate adjustment. Our study confirmed the previous finding that conditional on LTPA, relative OPA measure (RAS) that accounts for individual cardiorespiratory fitness was more predictive of AMI than absolute OPA measure (EE).[137,148] This finding is robust against the choice of OPA measure and modeling schemes, and agrees with an overall harmful effect of OPA on coronary heart disease (CHD) synthesized based on five prospective cohort studies published in 2011, 2012, and the first quarter of 2013[165] but not the other two prospective observational studies published earlier that suggest a protective effect[166] or no effect.[167] All seven studies simultaneously modeled OPA with LTPA: one found a positive association with IHD mortality;[139] one found a positive association with CHD incidence only among men with high LTPA level but not among men with low LTPA level;[140] three found no associations with MI[168] or CHD incidence,[167,169]; and two found negative

associations with IHD incidence when additionally accounting for occupational heavy lifting, or with CHD incidence when additionally accounting for commuting physical activity. Adjusted for LTPA, one recent study reported elevated risk for CHD mortality.[170] Additionally accounted for physical fitness, Holtermann and colleagues[171] found an increased risk of IHD mortality associated with high OPA in the least and moderately fit group, but not among the most fit men. Similarly, Clays et al.[142] reported a positive association between high OPA and mortality, and such association was particularly pronounced among workers with low physical fitness level. Our study used a unique RAS measure that took into account individual cardiorespiratory fitness, which, we believe, is crucial in studies examining the effect of different PA domains on cardiovascular health. The use of continuous versus broad categorical PA measures, difference in definition and categorization of these measures, and different study endpoints can be the reasons for the inconsistent findings in the current study as well as in the literature.[138]

Our study did not confirm the overwhelming evidence on and the long-held belief in the protective effect of LTPA on cardiovascular events.[165,172] For workers with high OPA, high LTPA was found to be preventive in some studies[139,168] but harmful in others.[140] A cluster randomized controlled trial also found opposing health impact of a 4-month aerobic exercise intervention among cleaners. While such intervention was found to increase cardiorespiratory fitness, lower resting and sleeping heart rate, and reduce relative workload (heart rate reserve), it also significantly increased systolic blood pressure.[173] These mixed findings cast doubt on whether the international physical

activity recommendation for the public at large[163] is similarly applicable to certain working populations with high physical job demands.

OPA and LTPA have different inherent characteristics. Mandatory OPA often has high frequency and long duration, involves activities such as heavy lifting, bending, pushing and pulling, monotonous and static postures with limited ability for pauses and restitution, and may not allow for adequate rest periods.[138,174] Such long-term high cardiovascular workloads can cause atherosclerosis via a prolonged elevated heart rate that leads to increased intravascular turbulence, unfavorable wall shear stress, endothelial injury,[175–178] and in turn inflammatory processes in the arterial walls,[137] according to the hemodynamic-inflammatory theory of atherosclerosis.[178] Other proposed physiological mechanism include high OPA-induced elevation in systolic blood pressure over the day (during work, at home and during sleep),[179] a strong predictor for cardiovascular events.[180] On the contrary, voluntary LTPA has shorter duration compared to OPA, involves more dynamic movements and sufficient variation and restitution.[174] Thus, people can achieve a training effect on the heart by performing relatively short (<1 hour/day) but intensive bursts of exhausting conditioning physical activity.[130]

When assessing the combined effect of both OPA and LTPA, we found men with high OPA and low LTPA had the highest risk for AMI. This is in line with the above theory on the different health impact of OPA versus LTPA, and the Belgian Physical Fitness Study.[142] However, another Belgian study showed an almost four times increased incidence of coronary events comparing men with high OPA and high LTPA to men with

low OPA and high LTPA.[140] A recent study in Israel found that employees who performed moderate-hard OPA (self-reported) and no LTPA had the greatest risk for all-cause mortality.[170] The combination of low OPA and low LTPA, averaging over levels of commuting PA, was associated with the highest risk for heart failure among Finnish men.[181] Given these mixed findings in the literature, the question of whether workers with high OPA should be advised to be highly physically active during leisure time,[174] remains open.

In our sensitivity analysis that used trichotomized PA measures, we divided the non-high PA category (originally labelled "low") into low and moderate PA categories. Among men without preexisting IHD, the highest AMI risk was observed among men with moderate OPA and moderate LTPA, followed by men with high OPA and low LTPA. Among men with IHD, the combination of low OPA and moderate LTPA, or high OPA and moderate LTPA, or moderate OPA and high LTPA positively predicted AMI when compared to the moderate OPA and low LTPA (lowest risk group). We did not choose to present these results as our main findings due to the somewhat arbitrary cutoff point for dividing low versus moderate PA and the relatively small sample size within each combination of OPA and LTPA. However, such exploration suggested interesting patterns that may be worth considering in future studies. First, the impact of OPA on AMI within a specific level of LTPA can be non-linear. As suggested by our analyses using continuous OPA measures, quadratic terms for PA measures may not capture such non-linearity. Second, how OPA is related to AMI (i.e. the shape of the relation) may differ across different levels of LTPA. These two points also apply to the relation

between LTPA and AMI by levels of OPA. Third, results from analyses that use dichotomized PA measures may depend on the cut-off points chosen for the categorization. To further complicate matters, the pattern seems to differ by preexisting IHD status as well.

Few studies examined the interplay between OPA, LTPA, and fitness on cardiovascular outcomes among workers with preexisting CVD. One study among Copenhagen men with preexisting CVD[182] found no association between moderate or high OPA and IHD mortality and a positive but uncertain association between high OPA and all-cause mortality. We observed greater impact of RAS on AMI at low compared to high LTPA level among men with IHD. Both studies failed to find an association between LTPA and cardiovascular outcomes. Compared to their counterparts free of IHD at baseline, these men had higher absolute EE and RAS levels but lower level of cardiorespiratory fitness. They may be more likely to experience an overloading associated with job-related heavy and especially static work on their cardiovascular system[183] and thus experienced a more detrimental health impact of high OPA than men without preexisting CVD. The positive but uncertain association between OPA and the outcome could be attributed to the small sample size. However, we cannot rule out the possibility that these employees who remained working, despite their preexisting conditions, were a selected, relatively healthy group.

Despite the fact that high RAS negatively predicted high LTPA in our sample (adjusted OR: 0.56, 95% CI: 0.44, 0.70 among men without IHD; adjusted OR: 0.51, 95% CI: 0.32,

0.83 among men with IHD), our hypothesized mediating pathway from OPA to AMI via LTPA was not supported due to the absence of an independent effect of LTPA on AMI after accounting for OPA and interaction between OPA and LTPA. In the current study, socioeconomic status as captured by education, cumulative measures for smoking and alcohol consumption, and psychosocial job factors are considered potential confounders and were adjusted for but biological factors were not. Different from PA measures that are reflective of their PA levels for the past year or even a longer period of time before baseline interview, biological measures such as blood pressure and blood glucose were assessed during baseline examination. Therefore, biological factors were conceptualized as mediating variables on the pathways from OPA or LTPA to AMI and not adjusted for in the main analyses. Analysis with additional adjustment for these factors showed that the positive OPA-AMI association at low LTPA attenuated, suggesting possible mediation by these factors. However, among men with IHD, positive OPA-AMI association persisted, despite the widened confidence interval. Future studies can examine the possible mediating role of both LTPA and these biological factors.[184,185]

The main strengths of our study include the prospective design, the representative sample of the population in Kuopio, long register-based follow-up, and adequate covariate adjustment. Also, the use of a validated detailed occupational interview produced better assessment of OPA compared to broad OPA categories obtained from most population-based surveys. The assessment of LTPA accounted for the seasonal variability of LTPA among Finnish men by averaging LTPA over a 12-month period. Different analytic strategies were implemented to disentangle the impact of OPA and LTPA on AMI. The

strategies included: accounting for cardiorespiratory fitness, using both categorical and continuous measures for OPA and LTPA, using a priori cutoff-points based on recommended and established guidelines for OPA and LTPA, and conducting several sensitivity analyses. Although previous studies have examined the interaction between OPA and LTPA, this is, to our best knowledge, the first study that used causal mediation analysis to examine the possible mediating pathway from OPA to AMI via LTPA.

Several limitations need to be addressed. Misclassification of OPA may be possible due to self-reporting rather than direct observations on type and duration of work activities and because energy expenditure assessment did not include upper extremity work or the handling of external loads. The EE also did not account for the amount of static work and ambient temperature, leading to a possibly conservative measure of the actual amount of energy expended at work.[137] Also, due to the lack of repeated cardiorespiratory fitness assessment at 4 years, we cannot compute a repeat RAS measure and further examine the change in RAS over time, leading to possible exposure misclassification. Examination of repeated measures for absolute EE in this cohort revealed relatively high correlation between baseline and 4-year absolute EE (correlation coefficient: 0.77). Due to the lack of 4-year RAS, our mediation analysis for estimating the path-specific effect of OPA on AMI via 4-year LTPA could be subject to intermediate confounding. Future work will involve conducting sensitivity analysis to check these results against the presence of such uncontrolled OPA-induced LTPA-to-AMI confounding that can introduce collider-stratification bias in the OPA-to-AMI relation.

In conclusion, our study contributed to clarifying the perplexing yet unsettled inter-relations among OPA, LTPA, and fitness in predicting AMI with and without preexisting conditions. Our finding on the positive association between OPA and AMI reaffirmed the need to develop physical activity recommendations that distinguish between OPA and LTPA,[186] and take into account individual worker health status, aerobic fitness, and physical demands of the job when designing strategies for CVD prevention for working populations.[138]

5.6 Appendix

Description of inverse-probability weighted (IPW) fitting of marginal structural models (MSM) for causal mediation analysis applied in the main analysis

Let X_0, M_0, and $\mathbf{Z}$ denote baseline OPA, baseline LTPA, and the set of covariates sufficient for confounding control. For binary OPA and LTPA, the steps are as follows. First, we created two copies of the original data set and included an additional variable ${X_0}^*$. ${X_0}^*$ was set to the actual value of OPA (i.e., ${X_0}^* = X_0$) for the first copy and was set to the opposite of the actual value of OPA (i.e., ${X_0}^* = 1 - X_0$). Then, in order to achieve both confounding control and effect decomposition, two sets of weights were computed: W_{X_0} and W_{M_0}. We modeled baseline OPA as a function of covariates and modeled baseline LTPA as a function of baseline OPA and covariates. The weight for each individual was calculated as:

$W = W_{X_0} \cdot W_{M_0}$ where $W_{X_0} = P(X_0 = x_0)/P(X_0 = x_0|\mathbf{Z} = \mathbf{z})$ and $W_{M_0} = P(M_0 = m_0|X_0 = {x_0}^*, \mathbf{Z} = \mathbf{z})/P(M_0 = m_0|X_0 = x_0, \mathbf{Z} = \mathbf{z})$. Finally, we ran a marginal structural Cox model (MSCM), weighted by W, on X_0, ${X_0}^*$, and $X_0 \cdot {X_0}^*$. The exponenttiated coefficient for X_0 was taken as the point estimate for PDE whereas the exponent of the linear combination of coefficients for both ${X_0}^*$, and $X_0 \cdot {X_0}^*$ was taken as the point estimate for TIE. Bias-corrected and accelerated 95% CIs were obtained based on 1000 bootstrap samples randomly selected from the original data with replacement.

To avoid unstable estimates due to extreme values of W_{X_0}, we also ran (1) a conditional MSCM, weighted by W_{M_0} and with adjustment for covariates, and (2) a doubly robust (DR) MSCM, weighted by W and with adjustment for covariates. The latter has DR property because we would obtain unbiased estimates for PDE and TIE as long as either the exposure model or the

final Cox model for the outcome was correctly specified. For continuous exposure and mediator, this approach became less ideal because this method requires substituting the probabilities [$P(X_0 = x_0|\mathbf{Z} = \mathbf{z})$ and $P(M_0 = m_0|X_0 = x_0, \mathbf{Z} = \mathbf{z}]$ in the weights by probability densities, which may yield unstable weights.[144] In the current study, only binary OPA and LTPA were examined in the mediation context. We further assumed that our models for the exposure and the mediator were correctly specified.

Description for sensitivity analysis of mediation by LTPA

Sample restriction

Of the total cohort of 2682 participants, a sub-cohort of men (N=1229) was actively followed at 4 and 11 years by examinations and questionnaire. Only 1038 participated at 4-year follow-up, after excluding 191 men due to death (N=35), severe illness (N=12), migration (N=5), no address (N=2), refusal (N=107), no contact (N=29) and other reason (N=1). We further excluded men who had retired (N=486) or had first-time AMI incidence after baseline (N=5) before 4-year follow-up, or had first-time AMI incidence within 1 year after the 4-year follow-up (N=1), or had missing data on key variables (N=41), leaving a sample of 505 men. Data involving 4-year follow-up on the sub-cohort were only used in the sensitivity analyses. Due to small number of men with IHD (N=50), we limited our analysis to men without IHD (N=455).

Method description

Due to the lack of 4-year repeated measurement of cardiorespiratory fitness, an important co-determinant of the health impact of OPA, mediating effect via 4-year OPA was not examined in the current study. With 4-year LTPA as an additional mediator, the PDE examined in the main

analysis where only baseline LTPA was considered as the mediator was further decomposed into the pathway effect of OPA on AMI that was through 4-year LTPA ($NIE_{4\text{-year}}$) only and the natural direct effect of OPA that was through neither baseline nor 4-year LTPA (NDE) (Figure A 5.1). To account for differential censoring by exposure, covariates and the outcome, we used inverse probability of censoring weights (IPCW) to reweight the sample in the final MSM so that censoring was statistically made to appear as a random event conditional on exposure, covariates, and the outcome.

Incidence of AMI for the restricted sample of 455 men without IHD

For the restricted sample of 455 men without preexisting IHD that has the start of follow-up as the survey date of 4-year follow up, 71 first-time incidence AMI occurred during an average of 16.96 years of follow-up (SD: 4.79; range: 1.01-20.81) and a total person-time of 7716 years.

Appendix Table A 5.5 depicted the results from this sensitivity analysis.

Appendix Tables: Results from sensitivity analyses

1. Use a trichotomized OPA measure and a trichotomized LTPA measure

1a. Effect estimates for one type of PA across levels of the other type of PA

Table A 5.1 Adjusted[a] hazard ratios (HR) and 95% confidence intervals (95% CI) for the main effect and joint effect of occupational physical activity (OPA) and leisure-time physical activity (LTPA) on 20-year incidence of acute myocardial infarction (N=495) when both types of physical activity were modeled as trichotomized variables, by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

		Men without IHD (N=1565)			Men with IHD (N=326)		
LTPA level	**OPA level**	**N**	**HR (95% CI)**	***P* for interaction[b]**	**N**	**HR (95% CI)**	***P* for interaction[b]**
Low (LTPA<20 minutes/week)	Low	63	Reference	--	5	Reference	--
	Moderate	111	1.11 (0.53, 2.29)	--	21	0.27 (0.07, 1.08)	--
	High	153	1.71 (0.88, 3.31)	--	54	0.56 (0.17, 1.87)	--
Moderate (20 minutes/week ≤LTPA<75 minutes/week)	Low	151	Reference	--	11	Reference	--
	Moderate	179	2.35 (1.43, 3.85)	0.093	30	0.39 (0.14, 1.11)	0.668
	High	166	1.91 (1.14, 3.20)	0.595	62	0.83 (0.33, 2.09)	0.011
High (LTPA≥75 minutes/week)	Low	323	Reference	--	28	Reference	
	Moderate	248	1.38 (0.94, 2.02)	0.796	52	2.31 (0.97, 5.48)	0.614
	High	171	1.31 (0.85, 2.02)	0.509	63	1.84 (0.77, 4.37)	0.118
OPA level	**LTPA level**						
Low (RAS≤23%)	Low	--	Reference	--		Reference	--
	Moderate	--	0.77 (0.37, 1.60)	--		1.04 (0.25, 4.37)	--
	High	--	0.99 (0.52, 1.91)	--		0.30 (0.07, 1.20)	--
Moderate (23%≤RAS<33%)	Low	--	Reference	--		Reference	--
	Moderate	--	1.64 (1.00, 2.70)	--		1.53 (0.55, 4.25)	--
	High	--	1.24 (0.75, 2.04)	--		2.58 (1.03, 6.45)	--
High (RAS>33%)	Low	--	Reference	--		Reference	--
	Moderate	--	0.86 (0.57, 1.30)	--		1.54 (0.89, 2.66)	--
	High	--	0.76 (0.50, 1.17)	--		0.98 (0.55, 1.72)	--

[a] Model adjusted for age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

[b] *P* values for the product term between the corresponding OPA and LTPA categories.

1b. Joint effect estimates of OPA and LTPA on AMI

Table A 5.2 Adjusted[a] hazard ratios (HR) and 95% confidence intervals (95% CI) for the main effect and joint effect of occupational physical activity (OPA) and leisure-time physical activity (LTPA) on 20-year incidence of acute myocardial infarction (N=495) when both types of physical activity were modeled as binary variables, by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

		Men without IHD (N=1565)[b]			**Men with IHD (N=326)[c]**		
LTPA level	**OPA level**	**N**	**HR (95% CI)**	**RERI (95% CI)[d]**	**N**	**HR (95% CI)**	**RERI (95% CI)**
Low	Low	63	1.29 (0.63, 2.67)	--	5	3.77 (0.93, 15.35)	--
(LTPA<20	Moderate	111	1.43 (0.79, 2.60)	-1.21 (-2.69, 0.27)	21	Reference	--
minutes/week)	High	153	2.21 (1.32, 3.69)	0.01 (-1.21, 1.23)	54	2.11 (0.85, 5.19)	--
Moderate	Low	151	Reference	--	11	3.92 (1.20, 12.78)	-0.38 (-6.05, 5.30)
(20 minutes/week	Moderate	179	2.35 (1.43, 3.85)	--	30	1.53 (0.55, 4.25)	--
≤LTPA<75							
minutes/week)	High	166	1.91 (1.14, 3.20)	--	62	3.24 (1.34, 7.83)	0.61 (-1.15, 2.36)
High	Low	323	1.28 (0.78, 2.13)	--	28	1.12 (0.36, 3.46)	-4.23 (-10.44, 1.98)
(LTPA≥75	Moderate	248	1.77 (1.08, 2.91)	-0.86 (-2.01, 0.29)	52	2.58 (1.03, 6.45)	--
minutes/week)	High	171	1.69 (1.00, 2.86)	-0.50 (-1.57, 0.56)	63	2.05 (0.83, 5.05)	-1.63 (-4.26, 1.00)

[a] Model adjusted for age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.
OPA levels: low (RAS≤23%), moderate (23%≤RAS<33%), high (RAS>33%)
[b] P=0.093 for the product term between moderate OPA and moderate LTPA. P values for other product terms are above 0.20.
[c] P=0.011 for the product term between high OPA and moderate LTPA. P=0.118 for the product term between high OPA and high LTPA. P values for other product terms are above 0.20.
[d] RERIs were measures for additive interaction and were calculated as $HR_{\text{index OPA, index LTPA}} - HR_{\text{index OPA, reference LTPA}} - HR_{\text{reference OPA, index LTPA}} + 1$.

2. Additionally adjust for biological factors

Table A 5.3 Adjusted[a] hazard ratios (HR) and 95% confidence intervals (95% CI) for the main effect and joint effect of occupational physical activity (OPA) and leisure-time physical activity (LTPA) on 20-year incidence of acute myocardial infarction (N=495) when both types of physical activity were modeled as binary variables, by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

	Men without IHD (N=1565) **HR (95% CI)**	**Men with IHD (N=326)** **HR (95% CI)**
Binary RAS (RAS>33% versus RAS≤33%)		
Low LTPA (< 75 minutes/week)	1.14 (0.86, 1.51)	1.43 (0.87, 2.33)
High LTPA (≥75 minutes/week)	0.91 (0.62, 1.33)	0.89 (0.51, 1.53)
Binary LTPA (LTPA≥75 minutes/week versus <75 minutes/week)		
Low OPA (RAS ≤ 33%)	1.02 (0.78, 1.34)	1.32 (0.76, 2.30)
High OPA (RAS > 33%)	0.81 (0.56, 1.19)	0.82 (0.51, 1.33)
P for Multiplicative Interaction[b]	0.281	0.421
Combination of RAS and LTPA, using low RAS and high LTPA as the reference		
Low OPA and high LTPA	Reference	Reference
Low OPA and low LTPA	0.91 (0.62, 1.33)	0.89 (0.51, 1.53)
High OPA and high LTPA	0.98 (0.75, 1.28)	0.76 (0.43, 1.31)
High OPA and low LTPA	1.12 (0.84, 1.50)	1.08 (0.66, 1.77)
RERI (95% CI)[c]	0.23 (-0.21, 0.67)	-0.58 (-1.58, 0.42)

[a] Model adjusted for age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, stress from work deadlines and biological factors including blood glucose, plasma fibrinogen, serum low-density lipoprotein cholesterol, serum high-density lipoprotein cholesterol, body mass index, systolic blood pressure, and taking lipid- or blood-pressure-lowering medication during follow-up as listed in Krause et al.[138]

[b] *P* value for the RAS × LTPA product term.

[c] RERIs were measures for additive interaction. For men without IHD, RERIs were calculated as $HR_{\text{high OPA, low LTPA}} - HR_{\text{high OPA, high LTPA}} - HR_{\text{low OPA, low LTPA}} + 1$. For men with IHD, RERIs were calculated as $HR_{\text{high OPA, high LTPA}} - HR_{\text{high OPA, low LTPA}} - HR_{\text{low OPA, high LTPA}} + 1$, using low OPA and low LTPA as the reference group.

3. Use continuous absolute energy expenditure (EE) as OPA measure

Table A 5.4 Hazard ratios (HR) and 95% confidence intervals (95% CI) for the main effect and joint effect of occupational physical activity (OPA), measured by continuous absolute energy expenditure, and continuous leisure-time physical activity (LTPA) on 20-year incidence of acute myocardial infarction (N=495), by preexisting ischemic heart disease (IHD) status, Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=1891).

	Men without IHD (N=1565)		**Men with IHD (N=326)**	
	Age-adjusted HR (95% CI)	**Fully-adjusted[a] HR (95% CI)**	**Age-adjusted HR (95% CI)**	**Fully-adjusted[a] HR (95% CI)**
Association between absolute EE (500 kcal/day increase) and AMI at levels of LTPA				
LTPA = 0 minute/week	1.04 (0.96, 1.12)	1.04 (0.96, 1.12)	0.95 (0.85, 1.07)	0.96 (0.86, 1.08)
LTPA = 75 minutes/week	1.07 (1.01, 1.13)	1.06 (1.00, 1.13)	0.98 (0.90, 1.08)	1.00 (0.91, 1.09)
Association between LTPA (75 minutes/week increase) and AMI at levels of absolute EE				
Absolute EE = 2111 kcal/day	0.95 (0.89, 1.03)	0.98 (0.91, 1.06)	0.95 (0.86, 1.06)	0.96 (0.86, 1.07)
Absolute EE = 2611 kcal/day	0.98 (0.90, 1.07)	1.01 (0.92, 1.09)	0.98 (0.87, 1.10)	0.99 (0.89, 1.11)
P for multiplicative interaction[b]	0.220	0.297	0.333	0.193

[a] Model adjusted for age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

[b] *P* value for the RAS × LTPA product term.

4. Sensitivity analysis for mediation of LTPA

Table A 5.5 Hazard ratios (HR) and 95% confidence intervals (95% CI) for the natural direct effect (NDE), natural indirect effect via baseline ($NIE_{Baseline}$) and 4-year ($NIE_{4\text{-}year}$) leisure-time physical activity of baseline occupational physical activity as measured by binary relative aerobic strain on 20-year incidence of acute myocardial infarction (N=71) among men without preexisting ischemic heart disease using inverse-probability weighted (IPW) fitting of marginal structural models (MSM)[a], Kuopio Ischemic Heart Disease Risk Factor Study, 1984-2011 (N=455).

Method	HR (95% CI)[b]
MSM[c]	
NDE	0.97 (0.48, 1.84)
$NIE_{baseline}$	0.79 (0.57, 1.10)
$NIE_{4\text{-}year}$	1.00 (0.87, 1.08)
Total effect	0.77 (0.41, 1.37)
Conditional MSM[d]	
NDE	1.22 (0.57, 2.46)
$NIE_{baseline}$	0.93 (0.69, 1.21)
$NIE_{4\text{-}year}$	0.96 (0.79, 1.05)
Total effect	1.08 (0.58, 2.02)
Doubly robust MSM[e]	
NDE	1.04 (0.45, 2.04)
$NIE_{baseline}$	0.92 (0.68, 1.20)
$NIE_{4\text{-}year}$	0.97 (0.82, 1.05)
Total effect	0.77 (0.41, 1.37)

[a] Occupational physical activity was measured by binary relative aerobic strain (RAS) indicator (RAS>33% versus RAS≤33%) and leisure-time physical activity was dichotomized (≥75 minutes/week versus <75 minutes/week). Covariates included age, education, participation in an unrelated clinical trial, smoking, alcohol consumption, mental strain at work, social support at work, and stress from work deadlines.

[b] Bias-corrected and accelerated 95% confidence intervals (CIs) were obtained using 1000 bootstrap samples.

[c] IPW was created based on a weight for OPA (dealing with confounding) and a weight for LTPA (decomposing effect).

[d] IPW was created based on a weight for LTPA (decomposing effect) only. Conditional MSM included covariates to control for confounding.

[e] IPW was created based on a weight for OPA (dealing with confounding) and a weight for LTPA (decomposing effect). In the final MSM, covariates were adjusted for.

Appendix Figures

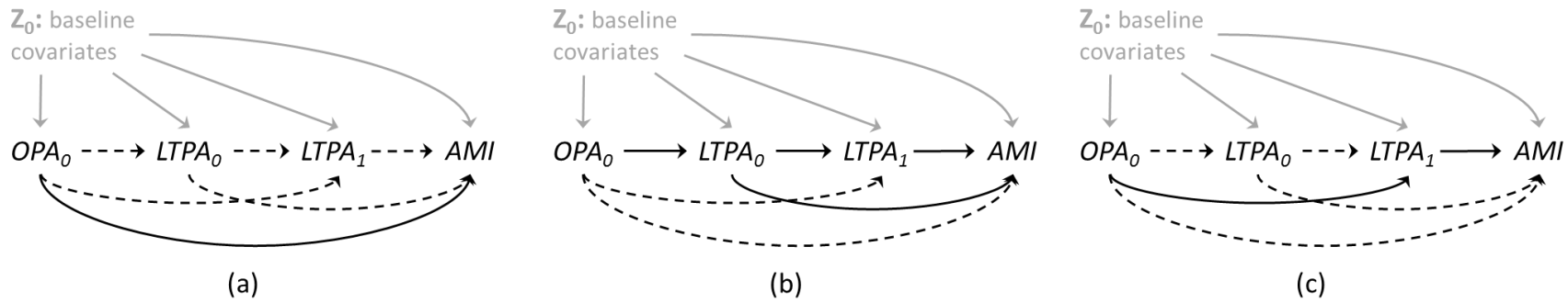

Figure A 5.1 Graphical presentation (solid black lines) of natural direct effect (a), natural indirect effect via baseline leisure-time physical activity (LTPA) (b), and natural indirect effect via 4-year LTPA only (c) of occupational physical activity (OPA) on acute myocardial infarction (AMI). Subscript 0 represents baseline measure and 1 represents 4-year measure.

Chapter 6. Concluding Remarks

From a methodological perspective, this dissertation provided a unified framework for estimating targeted effect(s) from the most recent 4-way decomposition in causal mediation analysis (single-mediator setting), and extended some of the existing estimation techniques to more complicated mediation settings. In Chapter 2, we first briefly reviewed the definition of different direct (natural and controlled), indirect, and interaction effects. Then we provided a three-step approach to implement g-computation via Monte Carlo simulation. This method provides an intuitive way of effect estimation by simulating and contrasting potential outcome values under different exposure intervention scenarios. We extended this framework to explore the mediating role of social factors and health behaviors in explaining education-related health disparities in Chapter 4. Effect estimation in this setting would be deemed almost impossible if we used other techniques not only because of examining multiple mediators simultaneously but also because the variables are of different types (continuous or categorical). G-computation, implemented via Monte Carlo simulation, is the most flexible approach and can be extend to longitudinal mediation analysis, despite its high demand of computation power.

Alternatively, depending on the variables involved and the data structure, other techniques can be used in causal mediation analysis such as regression-based technique[43] and inverse-probability-weighted (IPW) fitting of MSMs.[28,144] In Chapter 3, we extended the regression-based technique for single-mediator setting to a scenario involving contextual level exposure and intermediate confounding. Such extension was straightforward in a linear system and was easier to fit using existing software packages. In Chapter 5, we extended IPW of MSM method to a setting with time-varying mediator and survival outcome. Though theoretically possible, g-

computation applied to time-to-event data is very computationally intensive as it requires simulating conditional outcome (conditional on past exposure and survival history) at each unit of follow-up time.[55] The IPW method circumvents this by supplying individuals weights derived from models for exposure and mediator to achieve desired joint probability distribution, although it can be less efficient in some cases.

From an applied perspective, we explored the mechanisms underlying the effect of human development on individual health, the health disparity in education, and the effect of different physical activity domains on acute myocardial infarction, using various estimation techniques. Among low- and middle-income countries, we found a non-linear effect of human development on health, the majority of which was direct (not mediated by education nor BMI). Only a small portion of such impact was mediated by education and none was mediated by BMI. In the study of health disparity by education across countries over the world, we found that a substantial amount of the observed health disparities would be eliminated if everyone had healthy behaviors and achieving the desired level of social factors, although mechanistically, these factors were not the main contributor to the negative impact of lower education on health. In examining the interplay between occupational and leisure-time physical activity in affecting cardiovascular health, occupational physical activity was found to increase the risk of acute myocardial infarction at low but not high levels of leisure-time physical activity. Physical activity during leisure time did not mediate the impact of occupational physical activity on acute myocardial infarction. This dissertation has contributed to providing a clearer picture of how contextual as well as individual level causes affect health and offered some hypothesis and suggestions for further research in understanding the causal mechanisms at the end of each chapter.

In spite of the use of causal inference techniques, our results are based on the assumption of correct model specification for the outcome, mediator(s), or exposure (or some combinations of these), and the untestable assumption of conditional exchangeability (no uncontrolled confounding), as most observational studies do. No uncontrolled confounding assumptions required in mediation analysis are much stronger than those required for routine estimation of total effects. The presence of uncontrolled confounding for the mediator-outcome relationship given exposure and covariates is of particular concern. Reasons are the lack of measurements of such confounders in observational studies that are not designed to assess mediation, and no guarantee against such bias even when the exposure is perfectly randomized. We applied g-estimation technique to achieve statistical independence between the outcome (health) and the intermediate confounders of the mediator-outcome (BMI-health) relationship. In this way, the BMI-path-specific effect was estimated without bias (Chapter 3 appendix). This approach suggests that some of these estimation techniques can be used jointly to estimate path-specific effect in the presence of uncontrolled confounding bias, although effect decomposition becomes impossible. In the future, we will continue to examine the properties of each of these causal inference techniques in mediation analysis and try to incorporate bias analysis into the general mediation analysis framework by exploring and extending existing tools including an imputation-based technique for uncontrolled confounding of the total effect.[187]

As mentioned earlier in Chapter 2, causal mediation analysis has been criticized for its cross-world independence assumption[50] that can never be tested using data at hand. Nevertheless, effects defined in causal mediation analysis enjoy mathematical rigor and can still be useful in

disentangling mediation and interaction, and probing mechanistic questions. Estimation techniques presented in this dissertation such as g-computation can be used to estimate alternative causal parameters such as the stochastic controlled direct effect that does not suffer from this issue. When the satisfaction of certain assumptions is in doubt, researchers can present different mediation parameters to test competing hypothesis and conduct sensitivity analysis to check the robustness of the results to violations of key identification assumptions.

Incorporating causal diagrams and using counterfactual potential outcome framework, modern causal mediation analysis allows for nonlinearities and relaxes the no exposure-mediator interaction assumption imposed by traditional mediation analysis. This dissertation aims to advance the use of causal mediation analysis in probing mechanistic research questions by providing accessible tools for unpacking the black-box connection between exposure and health outcomes.

Reference

1. Hafeman DM, Schwartz S. Opening the Black Box: a motivation for the assessment of mediation. *Int J Epidemiol*. 2009;38(3):838-845. doi:10.1093/ije/dyn372.

2. Langer RD, Criqui MH, Reed DM. Lipoproteins and blood pressure as biological pathways for effect of moderate alcohol consumption on coronary heart disease. *Circulation*. 1992;85(3):910-915.

3. Suh I, Shaten BJ, Cutler JA, Kuller LH. Alcohol use and mortality from coronary heart disease: the role of high-density lipoprotein cholesterol. The Multiple Risk Factor Intervention Trial Research Group. *Ann Intern Med*. 1992;116(11):881-887.

4. Marmot MG, Elliott P, Shipley MJ, et al. Alcohol and blood pressure: the INTERSALT study. *BMJ*. 1994;308(6939):1263-1267. doi:10.1136/bmj.308.6939.1263.

5. Wright S. The Theory of Path Coefficients a Reply to Niles's Criticism. *Genetics*. 1923;8(3):239-255.

6. Wright S. The Method of Path Coefficients. *Ann Math Stat*. 1934;5(3):161-215.

7. Holland PW. Causal Inference , Path Analysis , and Recursive Structural Equations Models. *Sociol Methodol*. 1988;18:449-484.

8. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143-155.

9. Balke A, Pearl J. *Counterfactuals and Policy Analysis in Structural Models*. Morgan Kaufmann, San Francisco; 1995.

10. Halpern JY. Axiomatizing Causal Reasoning. *J Artif Intell Res*. 2000;12:317-337.

11. Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann; 2001:411-420. http://ftp.cs.ucla.edu/pub/stat_ser/R273-U.pdf.

12. Pearl J. Interpretation and identification of causal mediation. *Psychol Methods*. 2014;19(4):459-481. doi:10.1037/a0036434.

13. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173-1182.

14. Judd CM, Kenny DA. Process Analysis: Estimating Mediation in Treatment Evaluations. *Eval Rev*. 1981;5(5):602-619. doi:10.1177/0193841X8100500502.

15. James LR, Brett JM. Mediators, moderators, and tests for mediation. *J Appl Psychol*. 1984;69(2):307-321.

16. Glynn AN. The Product and Difference Fallacies for Indirect Effects. *Am J Pol Sci*. 2012;56(1):257-269. doi:10.1111/j.1540-5907.2011.00543.x.

17. MacKinnon DP, Lockwood CM, Brown CH, Wang W, Hoffman JM. The intermediate endpoint effect in logistic and probit regression. *Clin Trials*. 2007;4(5):499-513. doi:10.1177/1740774507083434.

18. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol*. 2007;58:593-614. doi:10.1146/annurev.psych.58.110405.085542.

19. Pearl J. The Mediation Formula : A guide to the assessment of causal pathways in nonlinear models Mediation : Direct and Indirect Effects. In: Berzuini C, Dawid P, Bernardinelli L, eds. *Causality: Statistical Perspectives and Applications*. Chichester, UK: John Wiley and Sons, Ltd; 2012:151-179.

20. Imai K, Keele L, Tingley D, Yamamoto T. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am Polit Sci Rev*. 2011;105(04):765-789. doi:10.1017/S0003055411000414.

21. VanderWeele T, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009;2:457-468.

22. Cole SR. Fallibility in estimating direct effects. *Int J Epidemiol*. 2002;31(1):163-165. doi:10.1093/ije/31.1.163.

23. Pearl J. Graphs, Causality, and Structural Equation Models. *Sociol Methods Res*. 1998;27(2):226-284. doi:10.1177/0049124198027002004.

24. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-688. doi:10.1093/biomet/82.4.669.

25. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37-48.

26. Rubin DB. Causal Inference Using Potential Outcomes. *J Am Stat Assoc*. 2005;100(469):322-331. doi:10.1198/016214504000001880.

27. Greenland S, Lash TL, Rothman KJ. Concept of Interaction. In: *Modern Epidemiology*. Third Ed. Lippincott Williams & Wilkins; 2008:71-86.

28. Vanderweele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*. 2014;25(2):300-306. doi:10.1097/EDE.0000000000000034.

29. Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics*. 2015;71(1):1-14. doi:10.1111/biom.12248.

30. Wang W, Nelson S, Albert JM. Estimation of causal mediation effects for a dichotomous

outcome in multiple-mediator models using the mediation formula. *Stat Med*. 2013. doi:10.1002/sim.5830.

31. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiol Method*. 2014;2(1):95-115. doi:10.1515/em-2012-0010.

32. VanderWeele T, Tchetgen ET. Mediation analysis with time-varying exposures and mediators. *Harvard Univ Biostat Work Pap Ser*. 2014. http://biostats.bepress.com/harvardbiostat/paper168. Accessed March 19, 2015.

33. Daniel RM, De Stavola BL, Cousens SN. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata J*. 2011;11(4):479-517.

34. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods*. 2013;18(2):137-150. doi:10.1037/a0031034.

35. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010;15(4):309-334. doi:10.1037/a0020761.

36. Jackson JW, VanderWeele TJ, Viswanathan A, Blacker D, Schneeweiss S. The explanatory role of stroke as a mediator of the mortality risk difference between older adults who initiate first- versus second-generation antipsychotic drugs. *Am J Epidemiol*. 2014;180(8):847-852. doi:10.1093/aje/kwu210.

37. Naimi AI, Moodie EEM, Auger N, Kaufman JS. Stochastic mediation contrasts in epidemiologic research: interpregnancy interval and the educational disparity in preterm delivery. *Am J Epidemiol*. 2014;180(4):436-445. doi:10.1093/aje/kwu138.

38. Zhang YT, Laraia B a., Mujahid MS, et al. Does food vendor density mediate the association between neighborhood deprivation and BMI? *Epidemiology*. 2015;26(3):344-352. doi:10.1097/EDE.0000000000000271.

39. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9-12):1393-1512. doi:10.1016/0270-0255(86)90088-6.

40. Robins JM., Hernan MA. Estimation of the causal effects of time-varying exposures. In: Verbeke G, Davidian M, Fitzmaurice G, Molenberghs G, eds. *Longitudinal Data Analysis*. Chapman and Hall/CRC; 2009:553-599. doi:10.1201/9781420011579.ch23.

41. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731-738. doi:10.1093/aje/kwq472.

42. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and

interactive effects. *Epidemiology*. 2013;24(2):224-232. doi:10.1097/EDE.0b013e318281a64e.

43. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition. *Epidemiology*. 2014;25(5):749-761. doi:10.1097/EDE.0000000000000121.

44. Vansteelandt S. Estimation of Direct and Indirect Effects. In: Berzuini C, Dawid P, Bernardinelli L, eds. *Causality: Statistical Perspectives and Applications*. Chichester, UK: John Wiley & Sons, Ltd; 2012:126-150. doi:10.1002/9781119945710.ch11.

45. Pearl J. The causal mediation formula--a guide to the assessment of pathways and mechanisms. *Prev Sci*. 2012;13(4):426-436. doi:10.1007/s11121-011-0270-1.

46. Rubin DB. Neyman (1923) and causal inference in experiments and observational studies. *Stat Sci*. 1990;5:472-480.

47. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.

48. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578-586. doi:10.1136/jech.2004.029496.

49. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010;21(4):540-551. doi:10.1097/EDE.0b013e3181df191c.

50. Richardson TS, Robins JM. *Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality*.; 2013. https://www.csss.washington.edu/Papers/wp128.pdf.

51. Naimi AI, Kaufman JS, MacLehose RF. Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects. *Int J Epidemiol*. 2014;43(5):1-6. doi:10.1093/ije/dyu107.

52. Tchetgen Tchetgen EJ, Vanderweele TJ. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*. 2014;25(2):282-291. doi:10.1097/EDE.0000000000000054.

53. Daniel R, Stavola B De, Cousens S, Leon D. Causal mediation analysis with multiple causally-ordered mediators. *Biometrics*. 2013;(January). doi:10.1111/biom.12248.

54. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne J a C. Methods for dealing with time-dependent confounding. *Stat Med*. 2013;32(9):1584-1618. doi:10.1002/sim.5686.

55. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology*. 2014;25(6):889-897. doi:10.1097/EDE.0000000000000160.

56. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci*. 2011;3(1):119-143. doi:10.1007/s12561-011-9040-7.

57. Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med*. 2012;31(18):2000-2009. doi:10.1002/sim.5316.

58. Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *Am J Epidemiol*. 2009;169(9):1140-1147. doi:10.1093/aje/kwp015.

59. Taubman SL, Robins JM, Mittleman M a, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009;38(6):1599-1611. doi:10.1093/ije/dyp192.

60. Cole SR, Richardson DB, Chu H, Naimi AI. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. *Am J Epidemiol*. 2013;177(9):989-996. doi:10.1093/aje/kws343.

61. Danaei G, Pan A, Hu FB, Hernán MA. Hypothetical midlife interventions in women and risk of type 2 diabetes. *Epidemiology*. 2013;24(1):122-128. doi:10.1097/EDE.0b013e318276c98a.

62. Garcia-Aymerich J, Varraso R, Danaei G, Camargo CA, Hernán MA. Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: an application of the parametric g-formula. *Am J Epidemiol*. 2014;179(1):20-26. doi:10.1093/aje/kwt229.

63. World Health Organization. World Health Survey: Guide to Administration and Question by Question Specifications. *Geneva*. 2002. http://www.who.int/healthinfo/survey/whsshortversionguide.pdf. Accessed July 16, 2013.

64. Simpson EH. The interpretation of interaction in contingency tables. *J R Stat Soc Ser B*. 1951;13(2):238-241. http://www.jstor.org/stable/pdf/2984065.pdf?acceptTC=true.

65. Vansteelandt S, Bekaert M, Lange T. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiol Method*. 2012;1(1):131-158.

66. Tchetgen EJT, Shpitser I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann Stat*. 2012;40(3):1816-1845.

67. Tchetgen EJT, Shpitser I. Semiparametric estimation of models for natural eirect and indirect effects. 2011.

68. van der Laan MJ, Petersen ML. Direct effect models. *Int J Biostat*. 2008;4(1). doi:10.2202/1557-4679.1064.

69. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*. 2009;20(1):18-26. doi:10.1097/EDE.0b013e31818f69ce.

70. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci*. 2010;25(1):51-71.

71. Albert JM. Distribution-free mediation analysis for nonlinear models with confounding. *Epidemiology*. 2012;23(6):879-888. doi:10.1097/EDE.0b013e31826c2bb9.

72. Hernán MA. Invited commentary: Agent-based models for causal inference—reweighting data and theory in epidemiology. *Am J Epidemiol*. 2015;181(2):103-105. doi:10.1093/aje/kwu272.

73. Wang A, Stronks K, Arah O a. Global educational disparities in the associations between body mass index and diabetes mellitus in 49 low-income and middle-income countries. *J Epidemiol Community Health*. 2014;68(8):705-711. doi:10.1136/jech-2013-203200.

74. Jamison DT. Investing in Health. In: Jamison DT, Breman JG, Measham AR, et al., eds. *Disease Control Priorities in Developing Countries.* 2nd editio. Washington, DC: Oxford University Press and The World Bank; 2006. http://www.ncbi.nlm.nih.gov/books/NBK11754/.

75. van der Kooi ALF, Stronks K, Thompson CA, DerSarkissian M, Arah OA. The modifying influence of country development on the effect of individual educational attainment on self-rated health. *Am J Public Health*. 2013;103(11):e49-e54. doi:10.2105/AJPH.2013.301593.

76. Rehkopf DH, Dow WH, Rosero-Bixby L. Differences in the association of cardiovascular risk factors with education: a comparison of Costa Rica (CRELES) and the USA (NHANES). *J Epidemiol Community Health*. 2010;64(9):821-828. doi:10.1136/jech.2009.086926.

77. Arah OA. On the relationship between individual and population health. *Med Health Care Philos*. 2009;12(3):235-244. doi:10.1007/s11019-008-9173-8.

78. James WPT, Jackson-leach R, Mhurchu CN, Kalamara E, Shayeghi M, Rigby NJ. Overweight and obesity (high body mass index). In: *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors.* In: Ezzati. Geneva, World Health Organization; 2004:497-596. http://apps.who.int/iris/bitstream/10665/42792/1/9241580348_eng_Volume1.pdf#page=521.

79. Wang A, Arah OA. Body Mass Index and Poor Self-Rated Health in 49 Low-Income and Middle-Income Countries, By Sex, 2002–2004. *Prev Chronic Dis*. 2015;12:150070. doi:10.5888/pcd12.150070.

80. Monteiro CA, Moura EC, Conde WL, Popkin BM. Socioeconomic status and obesity in

adult populations of developing countries : a review. *Bull World Health Organ*. 2004;82(12):940-946. http://www.who.int/bulletin/volumes/82/12/940.pdf.

81. WHO. World Health Survey. http://www.who.int/healthinfo/survey/en/. Accessed June 9, 2013.

82. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression , chronic diseases , and decrements in health : *Lancet*. 2007;370(9590):851-858. doi:10.1016/S0140-6736(07)61415-9.

83. Üstün TB, Chatterji S, Villanueva M, et al. WHO Multi-country survey study on health and responsiveness 2000–2001. In: Murray christopher J 0L., Evans DB, eds. *Health Systems Performance Assessment: Debates, Methods and Empericism.* Geneva: World Health Organization. http://www.who.int/healthinfo/survey/whspaper37.pdf.

84. Human Development Reports (HDR). *United Nations Dev Progr*. http://hdr.undp.org/en/. Accessed July 16, 2013.

85. Schlipköter U, Flahault a. Communicable diseases: Achievements and challenges for public health. *Public Health Rev*. 2010;32(1):90-119. http://www.scopus.com/inward/record.url?eid=2-s2.0-79959709098&partnerID=40&md5=c0e33d486afe342f6806807393b088f6.

86. Popkin BM. The nutrition transition and its health implications in lower-income countries. *Public Health Nutr*. 1998;1(01):5-21. doi:10.1079/PHN19980004.

87. Popkin BM, Adair LS, Ng SW. Global nutrition transition and the pandemic of obesity in developing countries. *Nutr Rev*. 2012;70(1):3-21. doi:10.1111/j.1753-4887.2011.00456.x.

88. Hawkes C, Chopra M, Friel S. Globalization, trade, and the nutrition transition. In: Labonté R, Schrecker T, Packer C, Runnels V, eds. *Globalization and Health: Pathways, Evidence and Policy.* New York: Routledge; 2009:235-262. http://www.globalhealthequity.ca/electronic library/Globalization, Trade, and the Nutrition Transition.pdf.

89. Schafer MH, Ferraro KF. Distal and variably proximal causes: education, obesity, and health. *Soc Sci Med*. 2011;73(9):1340-1348. doi:10.1016/j.socscimed.2011.08.010.

90. Adler NE, Stewart J. Health disparities across the lifespan: meaning, methods, and mechanisms. *Ann N Y Acad Sci*. 2010;1186:5-23. doi:10.1111/j.1749-6632.2009.05337.x.

91. Idler EL. Discussion: Gender Differences in Self-Rated Health, in Mortality, and in the Relationship Between the Two. *Gerontologist*. 2003;43(3):372-375. doi:10.1093/geront/43.3.372.

92. Patel V. Cultural factors and international epidemiology. *Br Med Bull*. 2001;57:33-45.

93. Rubin DB. Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *J Am Stat Assoc*. 1980;75(371):591-593.

94. Lahelma E, Manderbacka K, Rahkonen O, Karisto A. Comparisons of inequalities in health: evidence from national surveys in Finland, Norway and Sweden. *Soc Sci Med*. 1994;38(4):517-524.

95. Lahelma E, Martikainen P, Laaksonen M, Aittomäki A. Pathways between socioeconomic determinants of health. *J Epidemiol Community Health*. 2004;58(4):327-332.

96. Marmot MG, Stansfeld S, Patel C, et al. Health inequalities among British civil servants: the Whitehall II study. *Lancet*. 1991;337(8754):1387-1393. doi:10.1016/0140-6736(91)93068-K.

97. Kitagawa EM, Hauser PM. *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge, MA: Harvard University Press; 1973.

98. Pappas G, Queen S, Hadden W, Fisher G. The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. *N Engl J Med*. 1993;329(2):103-109. doi:10.1056/NEJM199307083290207.

99. Ross CE, Wu C. The Links Between Education and Health. *Am Sociol Rev*. 1995;60(5):719-745. http://www.jstor.org/stable/2096319.

100. Adler NE, Newman K. Socioeconomic Disparities In Health: Pathways And Policies. *Health Aff*. 2002;21(2):60-76. doi:10.1377/hlthaff.21.2.60.

101. Lalonde M. *A New Perspective on the Health of Canadians*. Ottawa; 1974. http://www.phac-aspc.gc.ca/ph-sp/pdf/perspect-eng.pdf.

102. Marmot MG, Kogevinas M, Elston MA. Social/economic status and disease. *Annu Rev Public Health*. 1987;8:111-135. doi:10.1146/annurev.pu.08.050187.000551.

103. Organization WH. A conceptual framework for action on the social determinants of health. 2010. http://www.who.int/iris/handle/10665/44489. Accessed September 8, 2015.

104. Feinstein E, Sabates R, Anderson TM, Sorhaindo A, Hammond C. What are the effects of education on health. In: *Proceedings of the Copenhagen Symposium*. OECD; 2006. https://www.researchgate.net/publication/242456417_4._What_are_the_effects_of_education_on_health.

105. US Department of Health E and W. *Healthy People: The Surgeon-General's Report on Health Promotion and Disease Prevention 1979*. Washington, DC; 1979. http://profiles.nlm.nih.gov/ps/access/NNBBGK.pdf.

106. Stringhini S, Dugravot A, Shipley M, et al. Health behaviours, socioeconomic status, and mortality: further analyses of the British Whitehall II and the French GAZEL prospective

cohorts. *PLoS Med*. 2011;8(2):e1000419. doi:10.1371/journal.pmed.1000419.

107. Lantz PM, House JS, Lepkowski JM, Williams DR, Mero RP, Chen J. Socioeconomic Factors, Health Behaviors, and Mortality. *JAMA*. 1998;279(21):1703. doi:10.1001/jama.279.21.1703.

108. Cutler DM, Lleras-Muney A. Understanding differences in health behaviors by education. *J Health Econ*. 2010;29(1):1-28. doi:10.1016/j.jhealeco.2009.10.003.

109. Marmot M, Wilkinson RG. Psychosocial and material pathways in the relation between income and health: a response to Lynch et al. *BMJ*. 2001;322(7296):1233-1236.

110. Cutler DM, Lleras-Muney A. Education and Health: Evaluating Theories and Evidence. 2006. http://www.nber.org/papers/w12352. Accessed August 4, 2013.

111. Sabia S, Elbaz A, Rouveau N, Brunner EJ, Kivimaki M, Singh-Manoux A. Cumulative associations between midlife health behaviors and physical functioning in early old age: a 17-year prospective cohort study. *J Am Geriatr Soc*. 2014;62(10):1860-1868. doi:10.1111/jgs.13071.

112. Institute of Medicine (US) Committee on Health and Behavior: Research; Practice; and Policy. Behavioral Risk Factors: The Interplay of Biological, Behavioral, and Societal Influences. 2001. http://www.ncbi.nlm.nih.gov/books/NBK43744/. Accessed September 16, 2015.

113. Jarvandi S, Yan Y, Schootman M. Income disparity and risk of death: the importance of health behaviors and other mediating factors. *PLoS One*. 2012;7(11):e49929. doi:10.1371/journal.pone.0049929.

114. Stringhini S, Sabia S, Shipley M, et al. Association of socioeconomic position with health behaviors and mortality. *JAMA*. 2010;303(12):1159-1166. doi:10.1001/jama.2010.297.

115. UN Common Database. United Nations Statistics Division. http://data.un.org/Data.aspx?d=WDI&f=Indicator_Code:NY.GDP.MKTP.CD. Accessed June 14, 2013.

116. Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. In: *Proceedings of the International Joint Conference on Artificial Intelligence*.; 2005:357-363.

117. Smith PM, Smith BT, Mustard CA, Lu H, Glazier RH. Estimating the direct and indirect pathways between education and diabetes incidence among Canadian men and women: a mediation analysis. *Ann Epidemiol*. 2013;23(3):143-149. doi:10.1016/j.annepidem.2012.12.012.

118. Marmot MG, Bosma H, Hemingway H, Brunner E, Stansfeld S. Contribution of job control and other risk factors to social variations in coronary heart disease incidence. *Lancet*. 1997;350(9073):235-239.

119. Braveman PA, Cubbin C, Egerter S, et al. Socioeconomic status in health research: one size does not fit all. *JAMA*. 2005;294(22):2879-2888. doi:10.1001/jama.294.22.2879.

120. Geyer S, Hemström O, Peter R, Vågerö D. Education, income, and occupational class cannot be used interchangeably in social epidemiology. Empirical evidence against a common practice. *J Epidemiol Community Health*. 2006;60(9):804-810. doi:10.1136/jech.2005.041319.

121. Adler NE, Rehkopf DH. U.S. disparities in health: descriptions, causes, and mechanisms. *Annu Rev Public Health*. 2008;29:235-252. doi:10.1146/annurev.publhealth.29.020907.090852.

122. Smith JP. Healthy bodies and thick wallets: the dual relation between health and economic status. *J Econ Perspect*. 1999;13(2):144-166.

123. United Nations. *Transitioning from MDGs to Post-2015 Development Agenda*.; 2015. http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG 2015 PR Transitioning.pdf.

124. Haapanen-Niemi N, Vuori I, Pasanen M. Public health burden of coronary heart disease risk factors among middle-aged and elderly men. *Prev Med (Baltim)*. 1999;28(4):343-348. doi:10.1006/pmed.1998.0426.

125. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Correction: actual causes of death in the United States, 2000. *JAMA*. 2005;293(3):293-294. doi:10.1001/jama.293.3.293.

126. Leitzmann MF, Park Y, Blair A, et al. Physical activity recommendations and decreased risk of mortality. *Arch Intern Med*. 2007;167(22):2453-2460. doi:10.1001/archinte.167.22.2453.

127. *Physical Activity Guidelines Advisory Committee Report*. Washington, DC; 2008. http://www.health.gov/paguidelines/report/. Accessed July 23, 2013.

128. Sofi F, Capalbo A, Cesari F, Abbate R, Gensini GF. Physical activity during leisure time and primary prevention of coronary heart disease: an updated meta-analysis of cohort studies. *Eur J Cardiovasc Prev Rehabil*. 2008;15(3):247-257. doi:10.1097/HJR.0b013e3282f232ac.

129. Nocon M, Hiemann T, Müller-Riemenschneider F, Thalau F, Roll S, Willich SN. Association of physical activity with all-cause and cardiovascular mortality: a systematic review and meta-analysis. *Eur J Cardiovasc Prev Rehabil*. 2008;15(3):239-246. doi:10.1097/HJR.0b013e3282f55e09.

130. Krause N. Physical activity and cardiovascular mortality – disentangling the roles of work, fitness, and leisure. *Scand J Work Env Heal*. 2010;36(5):349-355.

131. Barengo NC, Hu G, Lakka TA, Pekkarinen H, Nissinen A, Tuomilehto J. Low physical

activity as a predictor for total and cardiovascular disease mortality in middle-aged men and women in Finland. *Eur Heart J*. 2004;25(24):2204-2211. doi:10.1016/j.ehj.2004.10.009.

132. Hu G, Jousilahti P, Borodulin K, et al. Occupational, commuting and leisure-time physical activity in relation to coronary heart disease among middle-aged Finnish men and women. *Atherosclerosis*. 2007;194(2):490-497. doi:10.1016/j.atherosclerosis.2006.08.051.

133. Lopes C, Santos AC, Azevedo A, Maciel MJ, Barros H. Physical activity and risk of myocardial infarction after the fourth decade of life. *Rev Port Cardiol*. 2005;24(10):1191-1207.

134. Kannel WB, Belanger A, D'Agostino R, Israel I. Physical activity and physical demand on the job and risk of cardiovascular disease and death: the Framingham Study. *Am Heart J*. 1986;112(4):820-825.

135. Menotti A, Seccareccia F. Physical activity at work and job responsibility as risk factors for fatal coronary heart disease and other causes of death. *J Epidemiol Community Health*. 1985;39(4):325-329.

136. Kristal-Boneh E, Harari G, Melamed S, Froom P. Association of physical activity at work with mortality in Israeli industrial employees: the CORDIS study. *J Occup Environ Med*. 2000;42(2):127-135.

137. Krause N, Brand RJ, Kaplan GA, et al. Occupational physical activity, energy expenditure and 11-year progression of carotid atherosclerosis. *Scand J Work Env Heal*. 2007;33(6):405-424.

138. Krause N, Brand RJ, Arah OA, Kauhanen J. Occupational physical activity and 20-year incidence of acute myocardial infarction: results from the Kuopio Ischemic Heart Disease Risk Factor Study. *Scand J Work Environ Health*. 2014. doi:10.5271/sjweh.3476.

139. Holtermann A, Mortensen OS, Burr H, Søgaard K, Gyntelberg F, Suadicani P. The interplay between physical activity at work and during leisure time - Risk of ischemic heart disease and all-cause mortality in middle-aged Caucasian men. *Scand J Work Environ Heal*. 2009;35(6):466-474. doi:10.5271/sjweh.1357.

140. Clays E, De Bacquer D, Janssens H, et al. The association between leisure time physical activity and coronary heart disease among men with different physical work demands: A prospective cohort study. *Eur J Epidemiol*. 2013;28(3):241-247. doi:10.1007/s10654-013-9764-4.

141. Holtermann A, Marott JL, Gyntelberg F, et al. Does the benefit on survival from leisure time physical activity depend on physical activity at work? A prospective cohort study. *PLoS One*. 2013;8(1):e54548. doi:10.1371/journal.pone.0054548.

142. Clays E, Lidegaard M, De Bacquer D, et al. The combined relationship of occupational

and leisure-time physical activity with all-cause mortality among men, accounting for physical fitness. *Am J Epidemiol*. 2014;179(5):559-566. doi:10.1093/aje/kwt294.

143. US Department of Health and Human Services. *Physical Activity and Health: A Report of the Surgeon General.* Atlanta (GA); 1996. http://www.cdc.gov/nccdphp/sgr/pdf/sgrfull.pdf.

144. Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. *Am J Epidemiol*. 2012;176(3):190-195. doi:10.1093/aje/kwr525.

145. Salonen JT. Is there a continuing need for longitudinal epidemiologic research? The Kuopio Ischaemic Heart Disease Risk Factor Study. *Ann Clin Res*. 1988;20(1-2):46-50.

146. Tunstall-Pedoe H, Kuulasmaa K, Amouyel P, Arveiler D, Rajakangas AM, Pajak A. Myocardial infarction and coronary deaths in the World Health Organization MONICA Project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. *Circulation*. 1994;90(1):583-612.

147. Tuomilehto J, Arstila M, Kaarsalo E, et al. Acute myocardial infarction (AMI) in Finland--baseline data from the FINMONICA AMI register in 1983-1985. *Eur Heart J*. 1992;13(5):577-587.

148. Krause N, Brand RJ, Arah OA, Kauhanen J. Occupational physical activity and 20-year incidence of acute myocardial infarction: results from the Kuopio Ischemic Heart Disease Risk Factor Study. *Scand J Work Environ Health*. 2015;41(2):124-139. doi:10.5271/sjweh.3476.

149. Jørgensen K. Permissible loads based on energy expenditure measurements. *Ergonomics*. 1985;28(1):365-369. doi:10.1080/00140138508963145.

150. Holtermann A, Mortensen OS, Burr H, Søgaard K, Gyntelberg F, Suadicani P. Physical work demands, hypertension status, and risk of ischemic heart disease and all-cause mortality in the Copenhagen Male Study. *Scand J Work Environ Health*. 2010;36(5):357-365.

151. Karpansalo M, Lakka TA, Manninen P, Kauhanen J, Rauramaa R, Salonen JT. Cardiorespiratory fitness and risk of disability pension: a prospective population based study in Finnish men. *Occup Environ Med*. 2003;60(10):765-769.

152. Lakka TA, Laukkanen JA, Rauramaa R, et al. Cardiorespiratory Fitness and the Progression of Carotid Atherosclerosis in Middle-Aged Men. *Ann Intern Med*. 2001;134:12-20.

153. Lakka TA, Venalainen JM, Rauramaa R, Salonen R, Tuomilehto J, Salonen JT. Relation of Leisure-Time Physical Activity and Cardiorespiratory Fitness to the Risk of Acute Myocardial Infarction in Men. *N Engl J Med*. 1994;330(22):1549-1554.

doi:10.1056/NEJM199406023302201.

154. Taylor HL, Jacobs DR, Schucker B, Knudsen J, Leon AS, Debacker G. A questionnaire for the assessment of leisure time physical activities. *J Chronic Dis*. 1978;31(12):741-755. doi:10.1016/0021-9681(78)90058-9.

155. Salonen JT, Lakka TA. Assessment of physical activity in population studies - validity and consistency of the methods in the Kuopio ischemic heart disease risk factor study. *Scand J Sport Sci*. 1987;9(3):89-95.

156. Lynch J, Krause N, Kaplan G, Salonen R, Salonen J. Workplace demands, economic reward, and progression of carotid atherosclerosis. *Circulation*. 1997;96(1):302-307.

157. Lynch J, Krause N, Kaplan G, Tuomilehto J, Salonen J. Workplace conditions, socioeconomic status, and the risk of mortality and acute myocardial infarction: the Kuopio Ischemic Heart Disease Risk Factor Study. *Am J Public Health*. 1997;87(4):307-325. doi:10.1002/9781118229439.ch23.

158. Rose GA, Blackburn H, Gillum RF, Prineas RJ. *Cardiovascular Survey Methods*. 2nd Ed. Geneva: World Health Organization; 1982. http://www.ncbi.nlm.nih.gov/pubmed/4972212. Accessed December 16, 2013.

159. Cox DR, Oakes D. *Analysis of Survival Data*. New York: Chapman & Hall/CRC; 1984.

160. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. http://www.loc.gov/catdir/enhancements/fy0743/2007036316-d.html.

161. Knol MJ, VanderWeele TJ, Groenwold RHH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol*. 2011;26(6):433-438. doi:10.1007/s10654-011-9554-9.

162. Ilmarinen J. *Ergonomics Guidelines and Problem Solving*. Elsevier; 2000. doi:10.1016/S1572-347X(00)80015-4.

163. WHO. Global recommendations on physical activity for health. 2011. http://www.who.int/dietphysicalactivity/physical-activity-recommendations-18-64years.pdf.

164. Wang A, Arah OA. G-computation demonstration in causal mediation analysis. *Eur J Epidemiol*. 2015;30(10):1119-1127. doi:10.1007/s10654-015-0100-z.

165. Li J, Loerbroks A, Angerer P. Physical activity and risk of cardiovascular disease: what does the new epidemiological evidence show? *Curr Opin Cardiol*. 2013;28(5):575-583. doi:10.1097/HCO.0b013e328364289c.

166. Hu G, Jousilahti P, Borodulin K, et al. Occupational, commuting and leisure-time physical

activity in relation to coronary heart disease among middle-aged Finnish men and women. *Atherosclerosis*. 2007;194(2):490-497. doi:10.1016/j.atherosclerosis.2006.08.051.

167. Folsom AR, Arnett DK, Hutchinson RG, Liao F, Clegg LX, Cooper LS. Physical activity and incidence of coronary heart disease in middle-aged women and men. *Med Sci Sports Exerc*. 1997;29(7):901-909.

168. Holtermann A, Marott JL, Gyntelberg F, et al. Occupational and leisure time physical activity: risk of all-cause mortality and myocardial infarction in the Copenhagen City Heart Study. A prospective cohort study. *BMJ Open*. 2012;2(1):e000556. doi:10.1136/bmjopen-2011-000556.

169. Huerta JM, Chirlaque M-D, Tormo M-J, et al. Physical activity and risk of cerebrovascular disease in the European Prospective Investigation into Cancer and Nutrition-Spain study. *Stroke*. 2013;44(1):111-118. doi:10.1161/STROKEAHA.112.670612.

170. Harari G, Green MS, Zelber-Sagi S. Combined association of occupational and leisure-time physical activity with all-cause and coronary heart disease mortality among a cohort of men followed-up for 22 years. *Occup Environ Med*. 2015. doi:10.1136/oemed-2014-102613.

171. Holtermann A, Mortensen OS, Burr H, Søgaard K, Gyntelberg F, Suadicani P. Physical demands at work, physical fitness, and 30-year ischaemic heart disease and all-cause mortality in the Copenhagen Male Study. *Scand J Work Environ Health*. 2010;36(5):357-365.

172. Li J, Siegrist J. Physical activity and risk of cardiovascular disease--a meta-analysis of prospective cohort studies. *Int J Environ Res Public Health*. 2012;9(2):391-407. doi:10.3390/ijerph9020391.

173. Korshøj M, Lidegaard M, Skotte JH, et al. Does aerobic exercise improve or impair cardiorespiratory fitness and health among cleaners? A cluster randomized controlled trial. *Scand J Work Environ Health*. 2015;41(2):140-152. doi:10.5271/sjweh.3475.

174. Holtermann A. Occupational and leisure-time physical activity and coronary heart disease. *Occup Environ Med*. 2015. doi:10.1136/oemed-2015-102933.

175. Willerson JT, Golino P, Eidt J, Yao SK, Buja LM. Potential usefulness of combined thromboxane A2 and serotonin receptor blockade for preventing the conversion from chronic to acute coronary artery disease syndromes. *Am J Cardiol*. 1990;66(16):48G - 53G.

176. Zarins CK, Giddens DP, Bharadvaj BK, Sottiurai VS, Mabon RF, Glagov S. Carotid bifurcation atherosclerosis. Quantitative correlation of plaque localization with flow velocity profiles and wall shear stress. *Circ Res*. 1983;53(4):502-514.

177. A W, Fries CC, Sabini AM, Sawyer PN. The significance of turbulence in hemic systems and in the distribution of the atherosclerotic lesion. *Surgery*. 1965;57:155-162.

178. Glagov S, Zarins C, Giddens DP, Ku DN. Hemodynamics and atherosclerosis. Insights and perspectives gained from studies of human arteries. *Arch Pathol Lab Med*. 1988;112(10):1018-1031.

179. Clays E, De Bacquer D, Van Herck K, De Backer G, Kittel F, Holtermann A. Occupational and leisure time physical activity in contrasting relation to ambulatory blood pressure. *BMC Public Health*. 2012;12:1002. doi:10.1186/1471-2458-12-1002.

180. Hansen TW, Kikuya M, Thijs L, et al. Prognostic superiority of daytime ambulatory over conventional blood pressure in four populations: a meta-analysis of 7,030 individuals. *J Hypertens*. 2007;25(8):1554-1564. doi:10.1097/HJH.0b013e3281c49da5.

181. Wang Y, Tuomilehto J, Jousilahti P, et al. Occupational, commuting, and leisure-time physical activity in relation to heart failure among finnish men and women. *J Am Coll Cardiol*. 2010;56(14):1140-1148. doi:10.1016/j.jacc.2010.05.035.

182. Holtermann A, Mortensen OS, Burr H, Søgaard K, Gyntelberg F, Suadicani P. Fitness, work, and leisure-time physical activity and ischaemic heart disease and all-cause mortality among men with pre-existing cardiovascular disease. *Scand J Work Environ Health*. 2010;36(5):366-372.

183. Ilmarinen J. Work and cardiovascular health: viewpoint of occupational physiology. *Ann Med*. 1989;21(3):209-214.

184. Steeves JA, Bassett DR, Thompson DL, Fitzhugh EC. Relationships of occupational and non-occupational physical activity to abdominal obesity. *Int J Obes (Lond)*. 2012;36(1):100-106. doi:10.1038/ijo.2011.50.

185. Oppert J-M, Thomas F, Charles M-A, Benetos A, Basdevant A, Simon C. Leisure-time and occupational physical activity in relation to cardiovascular risk factors and eating habits in French adults. *Public Health Nutr*. 2006;9(6):746-754.

186. Kukkonen-Harjula K. Physical activity and cardiovascular health—work and leisure differ. *Scand J Work Environ Health*. 2007;33(6):401-404.

187. Thompson CA. Multiple bias modeling in a multi-center epidemiologic study of endometrial cancer - Dissertations & Theses - ProQuest. 2013. http://search.proquest.com/dissertations/docview/1412729897/7CF891DD6F8743E3PQ/3?accountid=14512.