UNIVERSITY OF CALIFORNIA

Los Angeles

How to Apply Directed Acyclic Graphs
to Descriptive, Predictive, and Causal Inference Aims
in Epidemiology

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Epidemiology

by

Ranjana Nisha Wickramasekaran

2023

ABSTRACT OF THE DISSERTATION

How to Apply Directed Acyclic Graphs
to Descriptive, Predictive, and Causal Inference Aims
in Epidemiology

by

Ranjana Nisha Wickramasekaran
Doctor of Philosophy in Epidemiology
University of California, Los Angeles, 2023
Professor Roch A.K. Nianogo, Co-Chair
Professor Onyebuchi A. Arah, Co-Chair

Applied epidemiologists are required to not only address causal aims but descriptive and predictive aims as well. There is a lack of guidance on how to approach aims that are not obviously causal with the causal tools and methods that epidemiologists are often trained in. Directed Acyclic Graphs (DAGs) are used in epidemiology and clinical research to clarify assumptions and illustrate causal questions to inform study design and statistical analysis. However, there is little guidance on the use of DAGs outside of causal inference. This dissertation aims to address this gap by walking through the use of DAGs while navigating and adapting previously developed frameworks. In chapter 1, we provide the background and general approach of the dissertation. In chapters 2-4, we adapt an existing framework to provide guidance on the use of DAGs to address descriptive, predictive, and causal aims, respectively. We demonstrate the application of DAGs by working through an example aim using data from the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS) as used in *Causal Inference: What If*. Lastly, chapter 5 provides a brief discussion of the similarities and differences in addressing these types of aims. We found that the importance of the target population is prevalent in any type of study. Similarly, selection bias, information bias, and missing data issues can arise in any study whereas confounding may not be as much of a concern in descriptive and some predictive studies. DAGs are useful to communicate and address these uncertainties.

The dissertation of Ranjana Nisha Wickramasekaran is approved.

Chad J. Hazlett

Tony Y. Kuo

Roch A.K. Nianogo, Committee Co-Chair

Onyebuchi A. Arah, Committee Co-Chair

University of California, Los Angeles

2023

DEDICATION

To my family and friends who carried me through my darkest hours. I wouldn’t be where I am today without your love and support. Thank you for always being there to pull me out of my grief, and anxiety. I could never express with words how much it’s meant to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF BOXES

LIST OF ACRONYMS

ALC	Alcohol frequency
ATE	Average treatment effect in the total population
ATT	Average treatment effect among the treated
ATU	Average treatment effect among the untreated
CDC	Centers for Disease Control and Prevention
CI	Confidence interval
DAG	Directed acyclic graph
DIET1	Post-smoking cessation diet
HT0	Height at baseline
IPCW	Inverse probability of censoring weight
IPTCW	Combined IPTW*IPCW weight
IPTW	Inverse probability of treatment weight
ITT	Intent-to-treat
KG	Kilograms
LATE	Local average treatment effect
MTRY	Number of variables randomly sampled as candidates at each split
NCHS	National Center for Health Statistics
NHANES-I	National Health and Nutrition Examination I
NHEFS	NHANES I Epidemiologic Follow-up Study
PAdl	Physical activity in daily life at baseline
PAdl1	Physical activity in daily life at follow-up
PArec	Recreational physical activity at baseline
PArec1	Recreational physical activity at follow-up
PSU	Primary sampling unit
Qsmk	Smoking cessation status
RSME	Root mean square error
SMKint	Smoking intensity (cigarettes per day)
SMKyrs	Years of smoking
WT0	Weight at baseline (kilograms)
WG	Change in weight at follow-up (kilograms)
WGc=0	Change in weight among the uncensored population (kilograms)

ACKNOWLEDGEMENTS

I would like to start by acknowledging my advisor, Dr. Onyebuchi Arah. Thank you for being an incredibly supportive and patient mentor. Your guidance and expertise have made me a better epidemiologist. Thank you for taking a chance on me. It has been an honor and privilege to work with you. I am also immensely grateful to Dr. Roch Nianogo. Thank you for always supporting me in my goals. Your friendship and guidance mean so much to me and I would be lost without it. Thank you to Dr. Tony Kuo for always seeing my potential even when I didn’t. I’ve grown so much as a researcher and epidemiologist since I began working with you and I’m immensely grateful for your support. Thank you to Dr. Chad Hazlett for your knowledge and guidance throughout the writing process of my dissertation. Your insights and expertise proved invaluable. To the faculty and staff at UCLA, thank you for your support. A special thanks to Joy Miller, Dr. Marjorie Kagawa-Singer, Dr. Karin Michels, Dr. Moira Inkelas, Dr. Frank Sorvillo, and Dr. Lisa Smith. Your support and guidance throughout the years have meant a lot to me. Thank you

To my colleagues at the Los Angeles County Department of Public Health, thank you for being amazing and supportive. A special thanks to Mirna Ponce Jewell, Brenda Robles, Lori Fischbach, and Noel Barragan for being my sounding board and keeping me on track. I am forever grateful.

To my friends, thank you so much for being the light in the darkness. I couldn’t have made it through without you. To my besties, Amy Gordon, Elizabeth Matson, and Sara Earley, your unwavering love and belief in me have made me a better person. To Alexander Martos-Manzur and Jen Frehn, thank you for listening to me and pulling me out of my stress and anxiety over the years of this program. To my friends and classmates at UCLA, Sylvia Tangney, Michael Dantès, Drake Edgett, Fan Zhao, Sreymom Oy, Melissa Soohoo, Matthew Coates, and Hoisum

Nguyen thanks for the support over the years. To my online gaming community for basically never making me feel alone. To Joel and Barbara Hipps and my friends at the institute, thanks for keeping me sane. I seriously couldn't have made it this far without any of you.

Lastly, to my family, for the love, support, and inspiration, thank you. To my parents, you have shaped who I am and I could not be more grateful for the unconditional love I have known. You have always inspired me to reach higher and I'm so happy to be your daughter. To my siblings, thanks for always having my back. I always know that I can count on you to be in my corner and I really appreciate that. To my nephews, thanks for putting a smile on my face even when I was stressed. I would also like to thank Belky and Gloria for everything you have done for me. Muchas gracias por su apoyo. To my family members who passed away, I will always cherish the love and memories I have of you and the impact you've made on my life.

VITA

EDUCATION

2013
Masters of Public Health, *Epidemiology*
University of California, Los Angeles, CA, USA

2008
Bachelor of Science, *Economics;* Minor, *Pure Mathematics*
Bachelor of Arts, *Music*
Loyola Marymount University, Los Angeles, CA, USA

RESEARCH EXPERIENCE

2018-2023
Epidemiologist, Research & Evaluation
Division of Chronic Disease & Injury Prevention
Los Angeles County Department of Public Health
Los Angeles, CA, USA

2015-2018
Epidemiology Analyst, Research & Evaluation
Division of Chronic Disease & Injury Prevention
Los Angeles County Department of Public Health
Los Angeles, CA, USA

2014-2015
Research Analyst, Office of the Deputy Director
Division of Chronic Disease & Injury Prevention
Los Angeles County Department of Public Health
Los Angeles, CA, USA

2012-2014
Research Assistant, Office of Senior Health
Division of Chronic Disease & Injury Prevention
Los Angeles County Department of Public Health
Los Angeles, CA, USA

TEACHING EXPERIENCE

Winter 2021
Teaching Assistant
Computer Management and Analysis of Health Data Using SAS
University of California, Los Angeles, CA, USA

SELECT PEER-REVIEWED PUBLICATIONS

Barragan NC, **Wickramasekaran RN**, Sorvillo F, Smith LV, Kuo T. Current trends in syphilis mortality in the United States, 2015-2020. *Venereology* 2023, 2(2), 59-64.

Wickramasekaran RN, Robles B, Dewey G, Kuo T. Evaluating the potential health and revenue outcomes of a 100% healthy vending machine nutrition policy at a large agency in Los Angeles County, 2013-2015. *J Public Health Manag Pract*. 2018; 24(3):215-224.

Wickramasekaran RN, Jewell MP, Sorvillo F, Kuo T. The changing trends and profile of pneumocystosis mortality in the United States, 1999-2014. *Mycoses*. 2017; 60(9):607-615.

SELECT PRESENTATIONS

Wickramasekaran RN, Wood M, Kwan A, Robles B, Reyes M, Kuo T. Comparing attitudes towards sodium reduction strategies across three California State Universities in Los Angeles County, 2019. *American Public Health Association 2020 Annual Meeting and Expo:* VIRTUAL MEETING. [Poster] 2020 (Oct 26)

Robles B, **Wickramasekaran RN**. Food Sector Partnership Landscape Analysis Results. *Food Service Guidelines Collaborative Annual Meeting*. Washington, D.C. [Invited Speakers] 2018 (June 28)

Wickramasekaran RN, Jewell MP, Sorvillo F, Kuo T. The changing trends and profile of Pneumocystosis mortality in the United States, 1999-2013. (Abstract #: 0243). *2016 Epidemiology Congress of the Americas:* Miami, FL. [Poster] 2016 (Jun 21)

SELECT AWARDS/HONORS

UCLA Department of Epidemiology Opportunity Award (2018-2019 school-year) for returning students pursuing a masters or doctoral degree, based on academic merit and personal achievements, financial need, and dedication to public health, Department of Epidemiology, UCLA Fielding School of Public Health, Los Angeles, CA, 2019.

Honorable Mention, awarded to authors of the peer-reviewed publication “Evaluating the potential health and revenue outcomes of a 100% healthy vending machine nutrition policy at a large agency in Los Angeles County, 2013-2015”, awarded by the Department of Public Health Science Summit for meritorious contribution to the advancement of science, Los Angeles County Department of Public Health, 2018.

Chapter 1 Introduction

1.1 Introduction

Directed Acyclic Graphs (DAGs) are used in epidemiology and clinical research to clarify assumptions and illustrate causal questions to inform study design and statistical analysis.[1–4] DAGs are generally considered to be *causal* diagrams. Causal diagrams are visual tools used to depict these types of relationships. DAGs provide a method to visualize and check dependencies among variables for model specification.[2] As such, they are instrumental in assessing whether certain conditions for identifiability have been met. Additionally, DAGs allow researchers to investigate different scenarios through the manipulation of variables. Thus, simulating counterfactuals and interventions becomes possible to explore.[2] DAGs are adaptable and flexible due to their non-parametric nature. Their mathematical foundations have been previously described.[1,5,6] However, the use of DAGs in applied research remains low.[7,8] A recent study found that around 40% of respondents did not use DAGs, and the most common reason given was that they did not know how to use them.[7] Accessibility to relevant training resources may be a barrier to more widespread use.[7]

1.1.1 Basics of DAGs

Although the basics of DAG use and construction have been described elsewhere,[1–5,9,10] it is important to reiterate these basics to understand their application in non-causal and causal studies. **Figure 1.1** and **Table 1.1** provide examples and definitions of key concepts related to DAGs. Graphs are made up of *nodes* and *arcs* where the nodes represent variables, whereas the arcs represent the relationship between variables. **Figures 1.1a** and **1.1b** are *directed* graphs because the arcs contained in the graph are all single-headed arrows. When the arrow from X to A is turned around as in **Figure 1.1b,** the graph is no longer *acyclic* since there is a feedback loop from X→A→C→X. Only **Figures 1.1a** and **1.1c** are acyclic because they do not contain any feedback loops. Therefore, **Figures 1.1a** and **1.1c** are both DAGs. It is important to note that the strongest assumption you can make in a DAG is the lack of an arc between two nodes since it is essentially the only way to indicate magnitude in a basic DAG, given that it represents no direct association between the two nodes.

To assess identifiability and/or bias in DAGs, we look at the *paths* that are presented in the DAG from the exposure of interest to the outcome of interest. Paths are a list of nodes and arcs that are all unique, regardless of the direction of the arrows. By tracing out the paths on a DAG, we can identify nodes as mediators, colliders, confounders, or instrumental variables. For instance, in **Figure 1.1a**, X←A→C←B→Y is a path. It is a *closed path* since it has a *collider*, C, on the path. A collider is a node that has two arrows pointing into it. Thus, information stops transmitting from X to Y at C in this path. However, X→M→Y is an *open path* as well as a *directed path* since the arrows are all pointing in the same direction from X to Y. On this path, X is an *ancestor* of Y and a *parent* of M since X and M are *adjacent.* In contrast, M is a *child* of X while Y is a *descendant* of X. In this case, M is a *mediator* since it is a node on a directed path. If we are interested in the *total effect* of X on Y, we would <u>not</u> want to adjust for this mediator because it would remove some of the effect of interest given that it is on a *causal path* (i.e., a directed path). However, X←C→B→Y is a *biasing path* since the path is open and not directed. It can also be called a *confounding path* since it begins with the exposure of interest, X, and ends with an arrow going into the outcome of interest, Y. As such, both B and C are considered *confounders* of the effect of X on Y. To ensure conditional exchangeability of the exposed and unexposed, it would be important to control for these confounders to assess the total effect of X on Y. Thus, we would want to consider adjusting for either B, C, or both to assess the total effect of X on Y. However, because C is a collider on path X←A→C←B→Y and a confounder on path X←C→Y, we cannot close one path without opening the other unless we also adjust for either A or B. Another confounding path is X←A→C→Y however, this path is also a *back-door path* since it is a path from X to Y and begins with a parent of X, A. Thus, according to this DAG, A, B, and C are sufficient to control for confounding. However, controlling for either only A and C or only B and C is *minimally sufficient* since either of these subsets of the sufficient set is enough to control for confounding. If C were an unmeasured variable as it is in **Figure 1.1c**, there would be no sufficient set for adjustment when using traditional methods to assess the direct effect of X on Y. However, if instead, we evaluate the effect of Z on Y, X then becomes a collider on the path from Z→X←(C)→Y. The only open path in **Figure 1.1c** is Z→X→Y. As such, we would call Z an *instrumental variable* because: 1) Z only affects Y through X; 2) Z is associated with X; and 3) Z is independent of confounders A, C, and B. Thus, in this scenario, we can use Z as an instrumental variable to assess the effect of X on Y.[5,11]

1.1.2 Justification

As previously mentioned, several resources have been published that provide the basics of DAGs.[1–5,9,10] Furthermore, Ferguson et al have provided guidelines on the synthesis of evidence and the construction of DAGs to provide a systematic method of the development of DAGs for causal inference.[12] Although DAGs have become an increasingly popular method for causal inference, the use of DAGs in applied research remains low.[7] Specifically, the use and reporting of DAGs vary in applied health research.[8] Even when DAGs are reported in these studies, most fail to report how adjustment sets were derived for estimates that are provided, including those for the primary analysis of interest.[8] Confusion or disagreement on the rules and assumptions of

DAGs may have contributed to their limited uptake as a systematic method to construct models.[7,8] DAGs are often taught from a causal perspective; therefore, applying them in non-causal or non-traditional settings may seem daunting. For example, knowledge of presenting some of the concepts unique to descriptive and predictive aims in a DAG is lacking and therefore may have prevented their use in these types of studies. Thus, a lack of access to relevant training resources may present a barrier to the more widespread use of DAGs.[7] A DAG may not be the most appropriate diagram at every stage of model specification; however, it can be a useful visual tool to communicate assumptions, check identifiability criteria, and discern sources of potential bias. This dissertation aims to develop and demonstrate guidance on the fuller use of DAGs for descriptive, predictive, and causal inference aims, which are commonly used in applied epidemiologic research.

1.2 Overall and Specific Aims

DAGs are used in epidemiology and clinical research as a tool to clarify assumptions and illustrate causal questions to inform study design and statistical analysis. They are especially useful in assessing whether certain conditions for identifiability have been met. DAGs are adaptable and flexible due to their non-parametric nature. Despite their well-established mathematical foundation, DAGs remain underused in applied research, often serving as token conceptual devices or flags when used.[7] Confusion or poor knowledge of the rules and assumptions of DAGs may have contributed to the limited or superficial uptake of DAGs in empirical work. Training resources present a barrier to the more widespread use of DAGs even in causal studies.[7] As such, low knowledge may have prevented their use in some settings such as studies with descriptive or predictive aims. To address this gap, this dissertation aims to develop and demonstrate guidance on using DAGs more fully from the beginning to the end of causal, predictive, and descriptive studies in applied epidemiology.

Aim 1:

To formalize and demonstrate the use of augmented directed acyclic graphs (DAG) for the design, analysis, and interpretation of descriptive aims in epidemiology, with application to the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS).

Aim 2:

To formalize and demonstrate the use of augmented DAGs the design, analysis, and interpretation of predictive model, applied to the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS).

Aim 3:

To formalize and demonstrate the use of augmented directed acyclic graphs (DAG) for the design, analysis, and interpretation of causal inference aims in epidemiology, with application to the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS).

1.3 Methods: Approach & Data

1.3.1 General Approach

We intend to carry out the aims of this dissertation in two main steps. The first step is to map each step of the study type to the use of DAGs and in what way, with or without augmentation. DAGs can be used to track background knowledge and the evolution of the data-generating process, including changes induced by conducting the study. The steps for each type of study that we will be basing our process on can be seen in **Table 1.2**. The steps for descriptive studies were adapted from a framework for reporting in descriptive studies by Lesko et al.[13] Those for predictive studies were inspired by the steps for the development of predictive models published by Steyerberg and Vergouwe.[14,15] Those for causal studies outlined in **Table 1.2** were adapted from the causal roadmap developed by Petersen and van der Laan.[16] The second step is to demonstrate the evolution of DAGs throughout the research stages through empirical applications using existing data.

1.3.2 Data for application and demonstration

To illustrate the evolution of DAGs during the data-generating process, we intend to the following dataset for empirical application:

1.3.2.1 The National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS)

The National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS) was a joint effort between the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies as a follow-up to investigate the relationships between clinical, nutritional, and behavioral factors that were previously assessed in NHANES I.[17] Specifically, the NHEFS study was intended to measure subsequent morbidity, mortality, hospital utilization, and changes in certain risk factors, functional limitation, and institutionalization. It is a national longitudinal study with a cohort that includes persons 25-74 who completed a medical examination for NHANES I in 1971-1975 (n=14,407).[17] To date, four follow-up studies were conducted in 1982-1984, 1986, 1987, and 1992. The first wave was conducted over two years (1982-1984) where personal interviews with subjects or their proxies were conducted; pulse rate, weight, and blood pressure of surviving participants were measured; hospital and/or nursing home records of overnight stays were collected; and death certificates of the decedents were also collected.[17] Follow-up continued in 1986, 1987, and 1992 where similar study design and data collection procedures were implemented. A 30-minute computer-assisted telephone interview was conducted instead of a personal interview. Additionally, no physical measurements were taken. The 1986 NHEFS was conducted among respondents to the previous NHEFS who were 55-74 years of age at baseline and not already known to be deceased (n=3,980).[17] In contrast, the 1987 and 1992 NHEFS follow-ups were conducted among the entire NHEFS cohort who were not deceased (n = 11,750 and 11,195, respectively).[17] Participants were successfully traced at some point through the 1992 follow-up at a rate of 96 percent.[17] We used the version of the dataset used in *Causal Inference: What If* where the data is restricted to individuals with known sex, age, race, weight, height, education, alcohol use, intensity of smoking at baseline and follow-up and those who answered the medical history questionnaire at baseline (n=1,629).[4,18]

1.4 Differentiating causal, predictive, and descriptive aims

As we can see in **Table 1.2**, the first step of any study regardless of its aim is to define the research question. It is at this point that we can determine whether the question of interest is truly causal, predictive, or descriptive in nature. Pearl's ladder of causation describes three rungs for causation: 1) association – seeing or observing the relationship based on the data that currently exists; 2) intervention – doing or intervening on something to observe the effect on the outcome of interest; and 3) counterfactual – imagining what could happen if an intervention had or had not taken place.[19] **Figure 1.2** provides a visual representation of the general relationship and scope of causal, predictive, and descriptive aims with respect to each other. This figure shows that all studies have some descriptive component to them. Whether a study aims to be purely descriptive, both predictive and causal studies aim to describe something. In predictive studies, the aim is to describe future or potential outcomes based on data that is currently available. In causal studies, the aim is to describe the relationship between exposure and an outcome. The differences in these aims can be more clearly seen in the scope of the aim. Purely descriptive aims are broader and tend to establish that a connection exists. Predictive aims tend to be narrower in scope as predictors are used to increase the connection between two variables such that the outcome of interest can be more accurately predicted. In Pearl's ladder of causation, diagnostic predictive and descriptive aims remain on the first rung of causation.[19] However, prognostic predictive aims depend on the treatment strategy used to predict a future outcome and can cross into the second or even third rung of causation.[20] Causal aims, however, are much more path-specific and dependent on the effect of interest and require conditions of identifiability to be met for causal inference. Whether the question calls for investigating the total, direct, or indirect effect requires an appropriate selection of deconfounders to isolate that pathway of interest. As such, they can land on any rung of the ladder of causation.

Figure 1.3 demonstrates a potential way to determine how to differentiate causal, predictive, and descriptive aims. Understanding how predictive and causal aims differ is necessary for methodological considerations and interpretation of the results in order to prevent conflation.[21] Conflation can occur when a study with a predictive aim is interpreted causally or variables selected are based on the causal structure and the presence of confounders, or conversely, a study with a causal/etiologic aim selects variables on their ability to predict the outcome.[21] This is not to say that there are no instances in which the lines between causal and predictive aims are blurred. In fact, counterfactual prediction is such an instance where the aim is to predict a counterfactual event and as such, requires causal inference.[21,22] Similarly, some treatment strategies within the prognostic framework may also require the consideration of the causal structure when the treatment strategy itself cannot be seen in the observational data and are therefore hypothetical.[20,21]

As such, it is important to understand a causal effect and how it differs from an association. In Pearl's ladder of causation, the difference can be described as seeing versus doing.[19] We see an association but we need to do something to observe an *effect*. Taken one step further would require us to imagine how something could have been. This alternate outcome is considered hypothetical or counterfactual since it cannot be presently observed given the current data.[19,20,22] When an action results in a particular outcome, the action taken is referred to as a *cause,* and the outcome that results from this cause is the effect.[4] Specifically, this outcome can be referred to as a causal effect. In mathematical terms, if there is a binary exposure X, the effect of X on Y would be considered a causal effect if the outcome Y when x=1 did not equal Y when x=0. If two actions are related, but there is not enough information to establish their relationship, the relationship is considered *associational.* In general, associational relationships lack a specified direction or can be bidirectional. In contrast, causal effects tend to have a specific order from action to outcome.

<u>1.5</u> **<u>Descriptive Aims</u>**

Descriptive epidemiology is generally non-causal in nature. The importance of these types of studies is understated in most introductory epidemiology courses.[23] They provide information on the trends and distribution of health, disease, and potential exposures in the population.

Specifically, they are typically used to describe patterns of exposure status or disease occurrence according to certain characteristics such as time, geographic location, age, race/ethnicity, and socio-economic status. They are particularly useful when little is known about the epidemiology of a disease.[24] As such, they tend to be a part of the initial discovery of information about the relationship between two variables. These studies can either stand on their own or be a first step to larger studies.[24] Since descriptive epidemiology can help identify patterns, it can serve several purposes: to aid with hypothesis generation; to guide the targeting of public health interventions; to assess health disparities; and to detect emerging health threats.[25] Disease surveillance is a common application of descriptive epidemiology.[25] Identifying the variation in certain characteristics of the population can help pinpoint sources of exposure. A historic example of descriptive epidemiology that is often referenced in introductory epidemiology courses is the investigation conducted by John Snow during the 1854 cholera outbreak in London, England. More recently, the application of descriptive epidemiology was frequently used during the beginning of the COVID-19 pandemic to understand the spread of disease and the frequency of disease occurrence among the most vulnerable populations over time.

1.5.1 Descriptive or causal?

Descriptive studies are often viewed as hypothesis-generating mechanisms for causal analyses however, this is not their sole purpose. As a result, the line between descriptive and causal studies can often be confusing. This may be because of the focus on causal epidemiology in public health programs rather than formally introducing students to what descriptive epidemiology entails.[23] While descriptive studies are intended to characterize variations in disease occurrence concerning person, place, and time of the population, analytic studies or studies of causation are intended to test causal hypotheses.[24,25] As such, a descriptive question aims to quantify and characterize the variation of a health feature of the population.[13,25] Given their non-causal nature, descriptive studies tend to describe associational relationships rather than causal effects.[26] In a sense, descriptive epidemiology deals more with discovery through correlation whereas causal epidemiology deals with testing a hypothesis. As with causal epidemiology, whether a study is intended to be descriptive or not depends on the research question that is meant to be answered.[13,23,27] The questions that they answer are typically with regards to defining the person, place, and time of the population in which disease occurs.[25] Similarly, they can also identify potential exposures in a population.

1.5.2 Bias in descriptive studies

Descriptive studies, like causal studies, are subject to confounding, selection bias, and measurement error.[13,23,28] Stratification, restriction, and adjustment may still be required in descriptive studies to reduce bias.[23] However, unlike causal studies where controlling for confounding is essentially mandatory, it may not be necessary to control for confounding in descriptive studies unless the question calls for it.[13,23,27] In fact, adjusting for confounding in descriptive studies can be harmful and result in over-adjustment.[13,23,27] However, mitigating selection bias and measurement error is necessary to ensure the results are valid.[13,23] Similarly, as with all epidemiologic studies, missing data issues can arise, and understanding the nature of missingness is important to properly address it.[13,28] It is also important to assess the plausibility of the assumptions made based on the type of missingness.[29] Thus, DAGs can be a useful tool to identify potential confounding variables, types of missingness, and bias.

1.6 Predictive Aims

The availability of big data has brought a larger focus on predictive studies, especially among health scientists. Predictive epidemiology is generally viewed as non-causal; however, causal inference can be a special case of prediction.[14,20,22,30] Predictive studies use data to predict potential outcomes (future or counterfactual) or observe the likelihood of the presence of the outcome based on current conditions. Thus, prediction models can be used to predict individual risk or identify strong predictors for the outcome.[14,15] In public health, predictive epidemiology can be used for both diagnostic and prognostic purposes.[14,30] In diagnostic settings, researchers use an individual's characteristics and/or symptoms to determine whether or not they currently have a specific disease. In prognostic settings, researchers aim to determine an individual's future outcome. Additionally, it is not uncommon for public health researchers to wish to know how a treatment or intervention may have affected (counterfactual prediction) or could affect (prognostic prediction) a patient's or population's potential outcome. These types of studies can involve causal inference which may blur the lines between causal and predictive studies; however, there are differences in approach and considerations that need to be made.

1.6.1 Predictive or Causal?

Causal inference and prediction modeling are typically viewed as separate branches in epidemiology. However, as with other study designs, the type of prediction model is dependent on the question being asked. In most cases, researchers use prediction models to predict the outcome of a disease based on the available data. However, in public health, researchers and policy-makers often like to know what the potential outcome could be had an intervention or policy taken place in a specific population. These questions are counterfactual in nature and require assumptions to be made about the relationships of predictors. Additionally, if the question is aimed at predicting a future outcome, it is in the interest of the reader to maximize the d-connectedness of the exposure and the outcome. Currently, in epidemiology, prediction is used to predict the exposure as with propensity score methods or to predict the outcome with G-computation. Both of these methods have been accepted within the causal inference framework. However, clinical predictive studies include analytic methods such as machine learning and simulation. To maximize the d-connectedness of the exposure and outcome, it is important to know the underlying mechanisms and relationships with respect to the exposure and outcome. Additionally, minimizing the number of predictors in the model may be of interest to improve the interpretability of the model.

1.6.2 Transportability and Predictor Selection in Predictive Studies

Although predictive studies are not generally subject to confounding bias,[30] they can be subject to the issue of *transportability*.[31,32] In epidemiology, we say that an estimate is transportable if we can extrapolate estimates derived from one population to another target population.[30,33,34] In diagnostic predictive studies, the transportability of the estimate to various populations is of great concern to validly predict the risk of the outcome from one population to the next. As such, the transportability of an estimate can be influenced by the predictors that are selected for inclusion in the model.[32] Methods for transporting a model/estimate to a different target population exist and can require causal assumptions to be met.[33–35] DAGs may be a useful tool to select predictors that can increase the transportability of the model to other populations.[32]

1.7 Causal Aims

Aims that require causal inference are a primary concern in epidemiology.[4,30] Whether it is identifying the etiology of a disease or establishing the effectiveness of a treatment/intervention, epidemiologists rely on causation to improve public health. Relying solely on non-causal evidence to diagnose and treat public health disparities is highly impractical and as such, causation is often the basis of epidemiology in schools of public health. It is from this perspective that DAGs are introduced to epidemiology students. As a result, DAGs are rarely used outside of the causal framework, and their use in applied epidemiology is limited.[7] Nonetheless, even within the causal framework, DAGs have been underused regardless of their flexibility and ability to evolve and be augmented. For causal aims, identification of the causal effect is the highest priority and DAGs can be used to clarify the steps needed for identification.

1.7.1 Conditions of Identifiability

As mentioned previously, DAGs are a good tool to assess whether certain identifiability conditions have been met. There are seven conditions of identifiability for causal inference (**Table 1.3**).[4,36–38] The first states that all variables are *well-defined*. In other words, all variables are properly named and measured. For example, a time-dependent variable should be properly labeled to identify when it was collected or measured. The next condition states that there is *exchangeability* of the risk of the outcome among the exposed and unexposed had the unexposed been exposed counter to fact. Thus, *conditional exchangeability* states that the outcome risk of the exposed (or treated) can be replaced by the potential outcome risk of the unexposed if, counter to fact, they had been exposed, conditional on certain covariates (i.e., (de-)confounders). Conditional exchangeability is often a more reasonable condition in observational studies, whereas exchangeability can be attained in large randomized trials.[4,30] Another condition that is useful for causal inference is *positivity*. Stated simply, positivity states that exposure values (i.e., exposed and unexposed) must be possible for all groups of the confounders under study, such that exposure is not deterministically assigned within levels of the confounding set.[4,30] The fourth condition, *consistency*, states that if a person is observed to be exposed, their outcome would be no different had they been assigned to be exposed. In other words, the outcome of a treated individual would be no different had the researcher intervened to assign said treatment to that individual.[4,30] The fifth states that there is no *interference* such that an individual's risk of the potential outcome is not affected by another individual's outcome (or even exposure). As such, their outcomes are independent. If interference is possible, it is essential to use models that account for it. The sixth condition states that there are *no other sources of bias*. Specifically, there is no selection bias, measurement error, or fraud. Lastly, the seventh states that there is *no model misspecification.* In other words, the statistical model must sufficiently account for existing sources of bias and should not introduce new sources.

1.7.2 Measures of interest

There are typically three types of causal estimands of interest: 1) the average treatment effect in the total population (ATE); 2) the average treatment effect among the treated (ATT); and 3) the average treatment effect among the untreated (ATU).[39,40] The ATE aims to measure the effect of treatment on the outcome in the total population. The ATT measures the effect of treatment among those who were actually exposed to the treatment while the ATU measures that among those who were actually unexposed. As previously mentioned, public health researchers are often interested in evaluating the effect of policy and interventions on health outcomes. To assess these questions, these researchers are interested in the causal effect of the intervention of interest among the treated, in other words, either the ATE or the ATT. Other estimands of interest include intent-to-treat analysis (ITT) and the local average treatment effect (LATE).[39,40] These estimands are typically used in instrumental variable analysis. The ITT measures the effect of an instrumental variable on the outcome whereas the LATE measures the effect of the treatment on the outcome among the subgroup in which the instrumental variable affects the treatment.[39,40]

1.8 Application Background

To demonstrate the differences and use of DAGs in descriptive, predictive, and causal studies, we will address the research questions surrounding smoking cessation and weight gain in **Box 1.1**. Smoking is extremely harmful and can affect nearly every organ in the body.[41,42] Each day around 1,600 youths try smoking a cigarette for the first time.[41] Moreover, smoking kills nearly one in five Americans annually.[42] Thus, quitting smoking has great benefits and can reduce a person's risk for serious smoking-related disease.[43] In fact, many adults would like to quit smoking and for the first time in the United States, there are more former smokers than current smokers as of 2022.[43] However, a potential barrier to smoking cessation is the potential for weight gain.[44] This can occur because of the way that nicotine affects the body.[44] For example, nicotine reduces food cravings and speeds up the metabolism. As a result, often a withdrawal symptom is increased food cravings not only because of replacement of cigarette cravings with potentially unhealthy foods, but also because of an improvement in food enjoyment.[44,45] When accompanied with a decrease in the metabolic rate, the risk for weight gain is increased. For demonstration purposes, we examine the relationship between smoking cessation and weight gain from descriptive, predictive, and causal angles using the NHEFS data.

1.9 Dissertation structure

This dissertation is structured into five chapters. In the first chapter (general introduction) we gave an overview of the current gaps in the use of DAGs, an introduction to DAGs, differentiated causal, predictive, and descriptive aims, and the overall aims of this dissertation. Additionally, we introduced the general approach and methods that will be used to address this gap in the following chapters. The second chapter will describe and demonstrate how DAGs can also be useful to address descriptive aims. The third chapter will take a similar approach, using DAGs when developing a predictive model to address predictive aims. The fourth chapter will break down how to use DAGs when applying the causal roadmap to answer causal questions. Lastly, the fifth chapter will provide a brief discussion of our findings.

1.10 Appendix

Table 1.1: Key terms, definitions, and examples for directed acyclic graphs (DAGs)

Terms	Definitions	Example
nodes	variables in a graph	X, the exposure of interest, is a node on the graph
edges/arcs	connectors (typically arrows) that are used to depict the relationship between two nodes	A is connected to C by an edge depicted as an arrow going from A to C
ancestor	a variable X is an ancestor of Y if it affects Y either directly or indirectly	A is an ancestor of M (A→X→M)
parent	a variable X is an ancestor of Y if it affects Y directly	X is a parent of M (X→M)
descendant	a variable Y is a descendant of X if it is affected by X either directly or indirectly	M is a descendant of A (A→X→M)
child	a variable Y is a descendant of X if it is affected by X directly	M is a child of M (X→M)
directed	all arcs in the graph are depicted by single-headed arrows	Figures 1A, 1B, and 1C are directed graphs
acyclic	no feedback loops	Figures 1A and 1C are acyclic whereas Figure 1B is not acyclic because it contains a feedback loop (X→A→C→X)
path	a sequence of nodes and edges from one node to another with no repeated nodes or edges, regardless of the direction of the arrows	X←A→C←B→Y is a path from X to Y in Figure 1A
causal/directed path	all arrows along the path are head-to-tail (i.e., all flow in the same direction)	X→M→Y is a causal path
collider	nodes that have two arrows pointing into them when tracing out a path	C is a collider on the path X←A→C←B→Y in Figure 1A
unblocked/open path	a path without colliders	X←C→Y is an open path
blocked/closed path	a path that has a collider on it	X→A→C←B→Y is a closed path
sufficient	a set of variables S where the effect of X on Y is unbiased given S	S_1={A,B,C} is a sufficient set for adjustment to estimate the total effect
minimally sufficient	a sufficient set of variables S where no proper subset of S is sufficient to identify the effect of X on Y	S_2={A,C} and S_3={B,C} are minimally sufficient sets for adjustment to estimate the total effect of X on Y in Figure 1A
biasing path	an open path that is not directed from X to Y	X←C←B→Y is a biasing path of the effect of X on Y

Terms	**Definitions**	**Example**
confounding path	an open path that is not directed from X to Y which ends with an arrow into Y	X←C→Y is a biasing path of the effect of X on Y
confounder	nodes that are common causes of X and Y or intercept a confounding path	C is a confounder on the path X←C→Y
back-door path	a path that connects exposure and outcome that begins with a parent of X	X←A→C→Y is a back-door path in Figure 1A
back-door criterion	Identifies a set of variables S that are sufficient to block all back-door paths once conditioned on	C is sufficient to block the backdoor paths X←C→Y, X←A→C→Y and X←C←B→Y and A is sufficient to block the backdoor path X←A→[C]←B→Y
mediator	nodes that intercept causal paths	M is a mediator of X and Y on the path X→M→Y
d-separation	two variables are independent of one another (no open paths between two nodes)	A is d-separated from B
instrumental variable	nodes that are ancestors of the exposure, only affect the outcome through the exposure and share no common causes with the outcome	Z is an instrumental variable

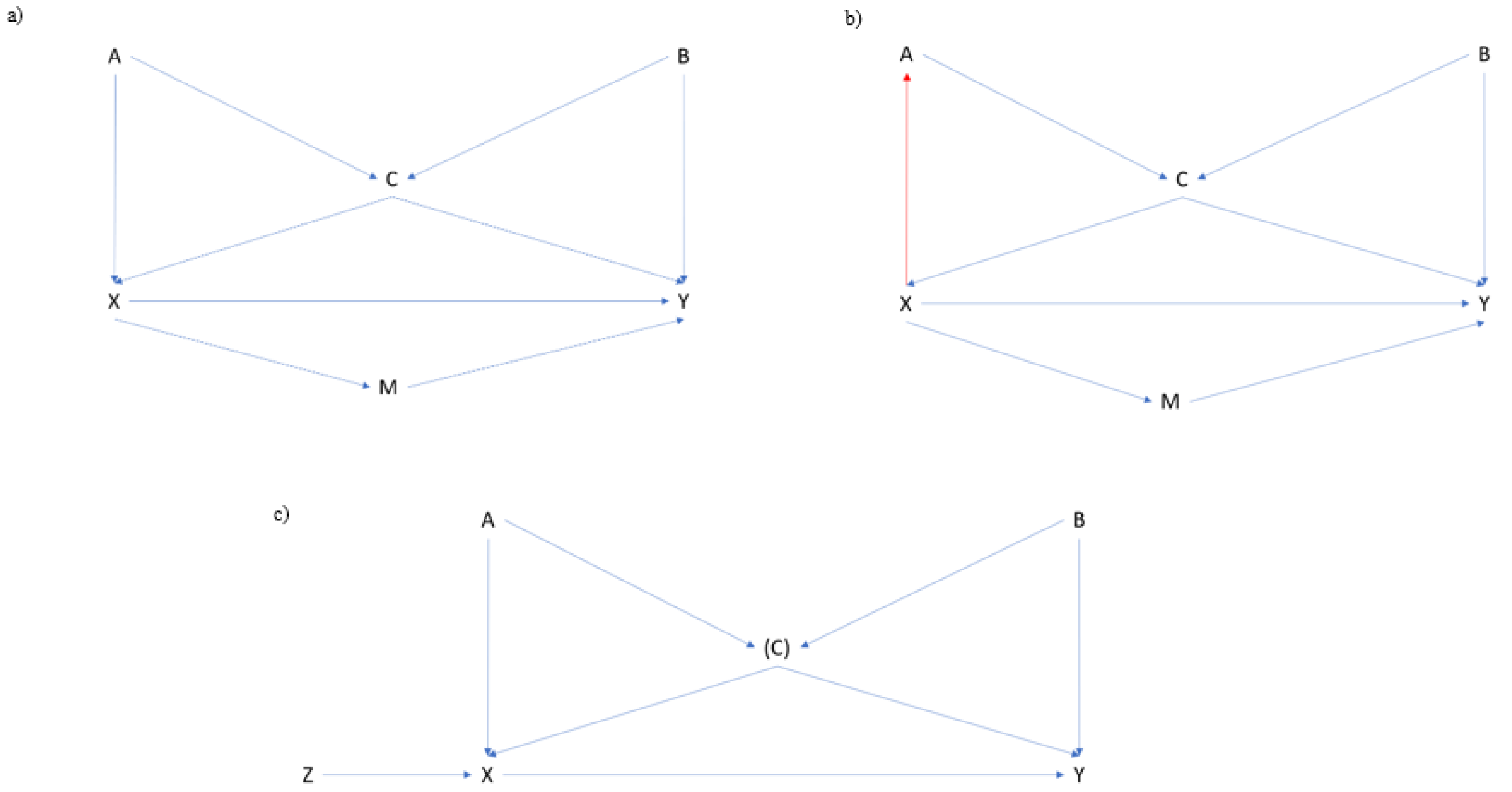

Figure 1.1: a) Example of basic DAG; b) example of a directed graph that contains a feedback loop; and c) example of instrumental variable DAG

Table 1.2: Comparing the steps of causal vs. predictive vs. descriptive studies

Descriptive	***Predictive***	***Causal***
0) Define the research question	0) Define the research question	0) Define the research question
1) Specify knowledge about the system	1) Specify knowledge about the system	1) Specify knowledge about the system
2) Link the data to the system	2) Link the observed data to the system	2) Specify the observed data
3) Specification of the target measure of occurrence & the role of covariates	3) Specification of the target quantity, model specification, and model estimation	3) Specification of the target causal quantity & assess identifiability
4) Estimation	4) Model performance & validation (Bias analysis)	4) Commit to a statistical model and estimand
5) Sensitivity/Bias analysis	5) Interpretation of results	5) Estimation
6) Interpretation		6) Sensitivity/Bias analysis
		7) Interpretation

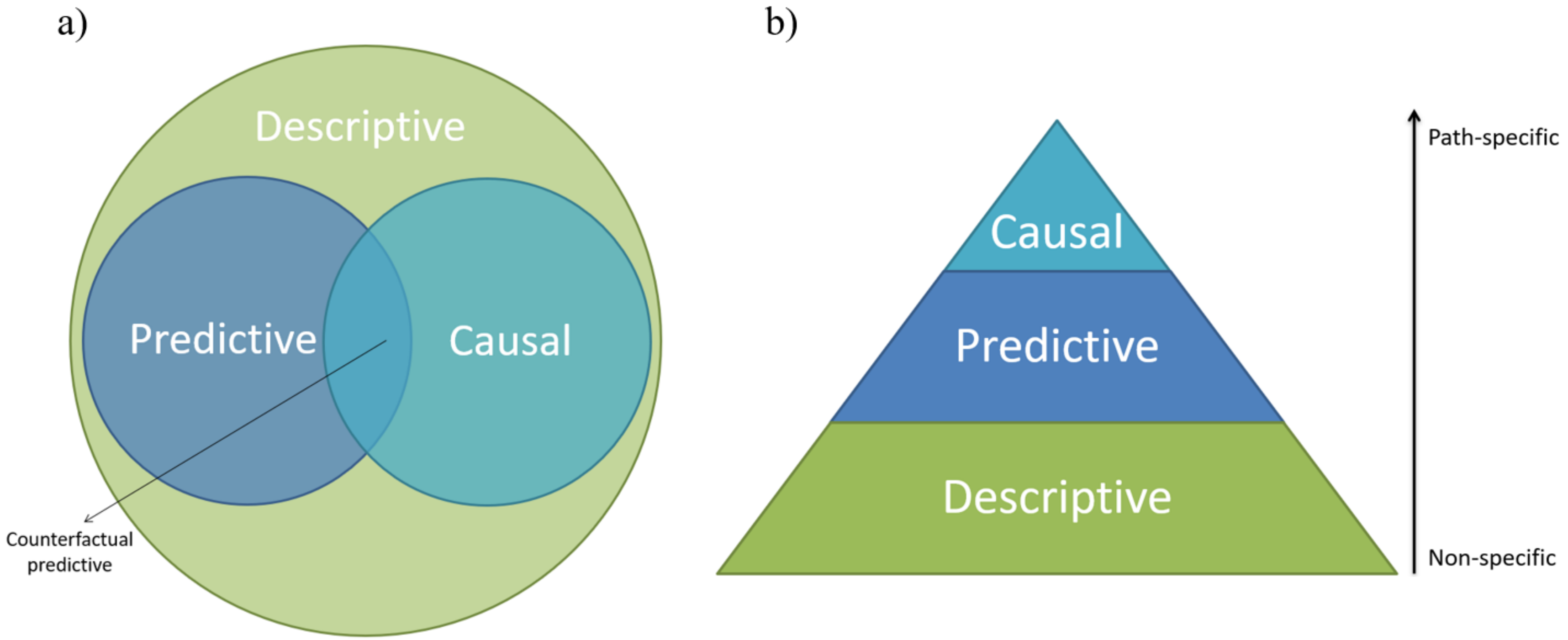

Figure 1.2: Illustration of the a) relationship and b) scope of descriptive, predictive, and causal aims

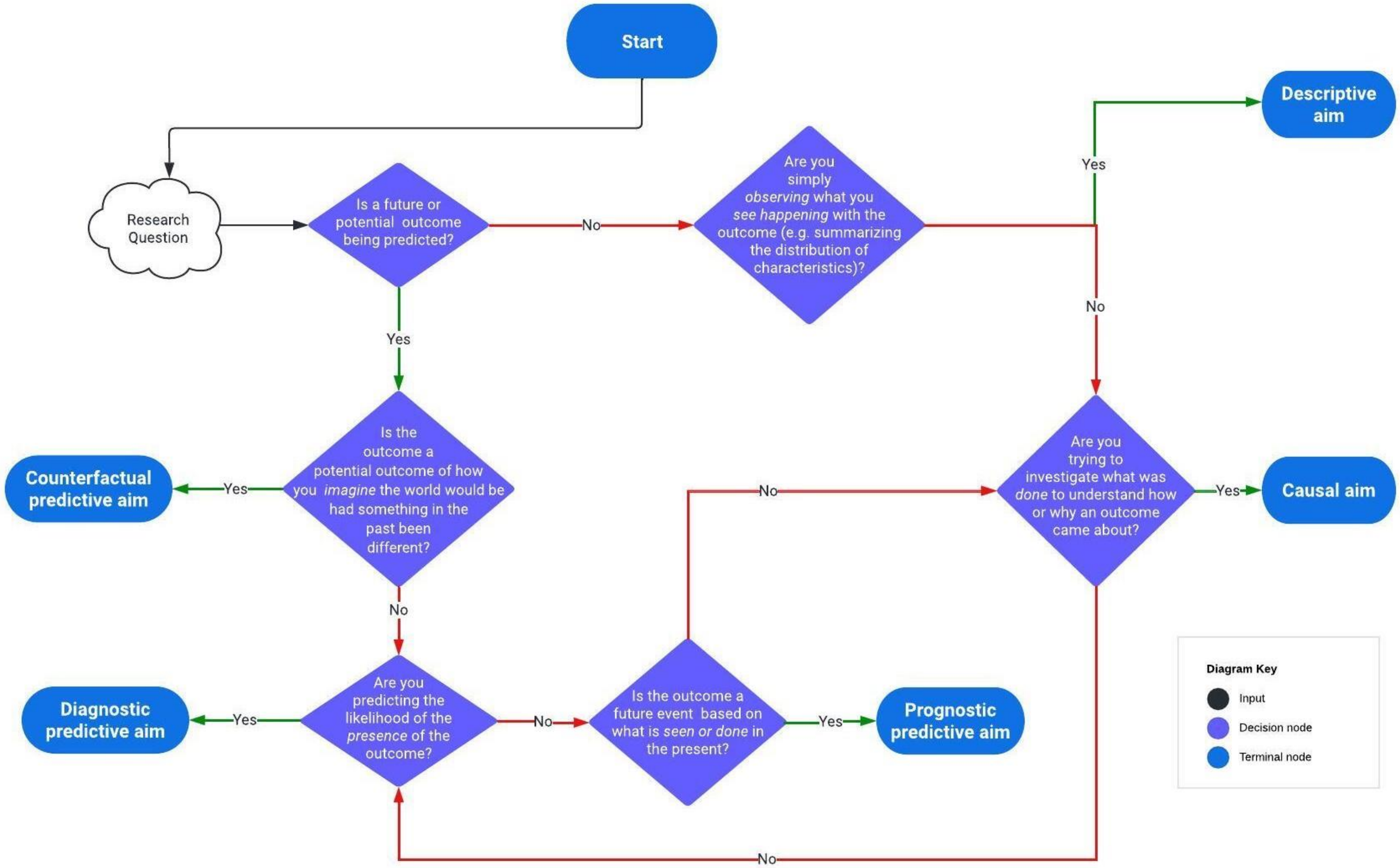
Start
Research Question
Is a future or potential outcome being predicted?
No
Are you simply *observing* what you *see happening* with the outcome (e.g. summarizing the distribution of characteristics)?
Yes
Descriptive aim
No
Yes
Is the outcome a potential outcome of how you *imagine* the world would be had something in the past been different?
Yes
Counterfactual predictive aim
No
Are you trying to investigate what was *done* to understand how or why an outcome came about?
Yes
Causal aim
No
Are you predicting the likelihood of the *presence* of the outcome?
Yes
Diagnostic predictive aim
No
Is the outcome a future event based on what is *seen or done* in the present?
Yes
Prognostic predictive aim
No
No
Diagram Key
Input
Decision node
Terminal node

Figure 1.3: Process map to determine the nature of the research question

Table 1.3: Conditions for effect identifiability

Condition	***Definition***
Well-defined variables	Every variable is properly named and measured
Conditional exchangeability	The potential outcome of *Y* had *X* been set to x is independent of *X* given a set of confounders *Z*
Positivity	For every non-zero probability of a set of confounders, *Z*, and exposure, *X*, there is a greater than zero probability of *X* given *Z* (i.e., all values of the exposure, X, must be possible for all Z under study)
Consistency	For those X = x, the potential outcome of Y had X = x is their observed Y
Interference	No spill-over effects or network ties between units. An individual's outcome is not dependent on another individual's outcome or exposure. Otherwise, if present, interference is appropriately accounted for.
No other sources of bias	No selection bias, dependent measurement error, or misrepresentation of the data
No model misspecification	The statistical model used sufficiently accounts for existing biases and no new biases are introduced

Box 1.1: Research questions for the application of descriptive, predictive, and causal aims

Descriptive - Who is gaining weight?

Predictive – What are the strongest predictors of weight gain?

Causal - What is the total average causal effect of smoking cessation on weight gain?

Chapter 2 DAGging out descriptive aims

2.1 Abstract

In epidemiology, researchers often conduct studies with descriptive aims to characterize patterns of disease, health, or even exposure to health risks in the population. Descriptive aims are especially helpful because they give context to person, place, and time of an event or disease of interest, allowing researchers to generate new hypotheses, assess health disparities, identify potential areas for targeted intervention, and describe the burden of disease or exposures to assess health needs. Here, we adapt previously introduced frameworks for reporting on descriptive aims[13] and the causal road map[16] to provide guidance on the use of directed acyclic graphs (DAGs) in the context of conducting descriptive studies. We then demonstrate one application of this adapted framework in an example analysis of weight gain in the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS). This chapter is intended to provide guidance on the incorporation of the use of DAGs in descriptive studies to assess and appropriately address sources of selection bias, misclassification bias, and missing data to support future research.

2.2 Introduction

In epidemiology, researchers often aim to describe patterns of disease, health, or even exposure to health risks in the population. Thus, the aim of the study is not necessarily causation but to establish correlation. In particular, studies with descriptive aims describe or document the distribution of disease or potential exposures in the population in terms of person, place, and time.[9,13,23–25,28] **Box 2.1** describes the person, place, and time of an event and the potential questions that may be of interest. Understanding the pattern of occurrence of disease in the population is a primary interest in public health research. Therefore, questions about whom the disease is affecting, where it is occurring, and how frequently the disease occurs are common. Addressing these questions serve several critical purposes in epidemiologic research (**Box 2.2**).[25] First, descriptive aims are commonly used to generate new hypotheses for analytic studies. These aims can help identify potential areas for more in-depth research. It is important to note that this is not the only potential purpose for descriptive aims. Second, descriptive aims can identify and assess health disparities in the population. Doing so can highlight potential exposures or systemic vulnerabilities that may have health consequences. This leads to the third purpose of descriptive aims which is to identify potential areas for targeted intervention. If one segment of the population is experiencing higher rates of disease, it is more cost-effective and urgent to concentrate public health efforts there. Lastly, descriptive aims have been used to detect emerging public health threats. An example of this last purpose is public health surveillance. Descriptive aims are commonly addressed because they serve a critical role in public health research.

Despite their importance and common application in public health research, epidemiologic studies with descriptive aims are rarely taught beyond introductory epidemiology classes.[23] Additionally, tools that are acquired in school, such as directed acyclic graphs (DAGs), are taught from a causal perspective. As a result, there is little guidance on how to apply DAGs to descriptive aims.[7] Here, we demonstrate the use of DAGs to address descriptive aims and expand on the descriptive framework provided by Lesko et al. We do this by reviewing, reasoning about, and annotating where and how DAGs could be useful in conducting a study with descriptive aims by adapting Lesko et al.'s descriptive framework and the causal roadmap by Petersen and van der Laan.[13,16] We illustrate these with application to data taken from *Causal Inference: What If* and addressing the question of who is gaining weight in the analytic sample used in the book.[4]

2.3 Applying DAGs to the descriptive framework

2.3.1 Overview of the descriptive framework

Lesko et al. developed a checklist for reporting on descriptive aims (**Table 2.1**).[13] Although this checklist is comprehensive for reporting, questions on how to utilize DAGs to inform study design, analysis, and interpretation for a descriptive aim remain.[7] We develop steps to conduct a descriptive study in alignment with this checklist and incorporate how DAGs can be used to address these items (**Table 2.2**). The steps we follow include: 0) specifying the research question; 1) specify knowledge about the system; 2) linking the data to the system under study; 3) specifying the measure of occurrence and the role of covariates; 4) conducting a sensitivity/bias analysis; 5) estimation; and 6) interpretation. In this paper, we apply DAGs to a descriptive aim and demonstrate their use with application.

2.3.2 Step 0: Specify the research question

A preliminary step of all studies is to specify the research question and studies with descriptive aims are no different. The importance of having a well-defined research question cannot be overstated.[27] A descriptive research question aims to strictly *see* what is happening in a population.[19,22,23,30] There are four main characteristics of a well-defined research question for descriptive aims (**Box 2.3**).[13,23] The first is defining the target population.[13,23] This is especially important for descriptive aims.[13,23,27] The research question should address who the target population is. A well-defined target population should describe who, where, and when the study is aiming to make inferences about. Platt argues that defining the target population may be an unnecessary step in some descriptive aims where the inference is about the sample itself.[28] This may be true, however, it may be beneficial to be transparent about this aim from the start to prevent others from overgeneralizing any results in the future. Second, the research question defines the outcome, event, or characteristic of interest. Along with this, the third characteristic of the research question is specifying the measure of occurrence of interest. Generally, descriptive research questions will aim to quantify the prevalence, incidence, or frequency of a certain outcome in the population.[13,23] This will be made clear in the research question itself. Lastly, there may be other variables that need to be specified as potential covariates. These variables will either be stratified on or standardized over depending on whether they are variables that further characterize the outcome or detract from the ability to quantify the outcome.[13,23]

2.3.2.1 Step 0: Application

To demonstrate the application of this framework with DAGs, we will look at who is gaining weight and assess crude mean and risk differences in weight gain on smoking cessation status.

This example is a descriptive aim to expand on an example presented by Hernán and Robins in Chapter 12 of *Causal Inference: What If.*[4] Specifically, we would like to know the proportion of individuals who gained a significant amount of weight from 1971 to 1982 overall, as well as among the crude mean or risk difference of weight gain among those who quit smoking compared to those who did not. The time period for this study is 1971 to 1982. The target population, in this case, is our analytic sample which consists of cigarette smokers ages 25-74 years who remain in the analytic sample after exclusion criteria is applied. Our outcome of interest is weight gain at the end of the study period. The aim is to describe who gained weight and quantify the differences between those who quit smoking and those who did not. We would also like to observe any differences in the frequency of weight gain for demographic categories and certain lifestyle characteristics (i.e., daily activity level, recreational activity level, and smoking intensity). Similarly, we would like to observe any demographic and characteristic differences between those who had clinically significant weight gain and those who did not.

2.3.3 Step 1: Specify knowledge about the system

For descriptive aims, it may not be the case that knowledge of the system exists. However, it may still be useful to illustrate exactly what we aim to assess and its limitations. For example, if we wish to characterize an outcome Y, we may not know the direction of the effect of the characteristic on Y or vice versa unless we have longitudinal data. However, with a descriptive aim, this may not be necessary. We may only be interested in quantifying the presence of the outcome during a certain time period. So why use DAGs? Why even have this step? Simply, descriptive aims can serve as more than a hypothesis-generating mechanism. They are population-specific and as such, the same descriptive question in a different population is still valuable information for discovery. For instance, they can also be used to allocate resources to populations in need. Additionally, it may be beneficial to outline what is already known to better identify critical information for data collection. If data has already been collected, they may provide a basis for narrowing the field of study. Thus, DAGs may be helpful to justify such choices since they are valuable tools for communication, even when the aim is non-causal. DAGs can communicate the assumptions and hypotheses of the researcher and as such are valuable tools for transparency of why variables were collected/assessed with the outcome.

2.3.3.1 Step 1: Application

Figure 2.1 shows potential DAGs for the relationships between baseline covariates and the outcome(s) of interest. In our example, our primary outcome of interest is weight gain at the end of the study period (*wg*). After a review of the literature, we constructed a DAG to illustrate the relationship between the variables of interest and the outcome (**Figure 2.1a**) as well as a simplified DAG where we assume that the relationships between baseline covariates are not of interest (**Figure 2.1b**). For the remainder of the study, we will work with the DAG in **Figure 2.1b** for simplicity. In this DAG, Z_0 refers to the baseline covariates age, education, sex, race, recreational physical activity, daily life physical activity, smoking intensity, years of smoking, and weight at baseline. The mediators (M_1) are recreational physical activity, daily life physical activity, and diet between baseline and follow-up as a result of smoking cessation. Lastly, the main covariate of interest is smoking cessation status (*qsmk*) between baseline and follow-up. Though smoking cessation status is technically time-varying, we will ignore this as Hernán and Robins do in Chapter 12 of *What If*.[4] Our primary interest is to quantify the distribution of demographic and behavioral characteristics with respect to weight gain.

2.3.4 Step 2: Link the data to the system

In this step, we aim to assess the internal validity and potential external validity of the analysis given the data. When addressing a descriptive aim, there is generally less concern about confounding.[23] However, selection bias and measurement error may still be a concern. People may be differentially selected for the study such that the data reflects only part of the story. Additionally, the quality of the classification of the outcome needs to be investigated to ensure that misclassification bias is mitigated. Similarly, missing data is a potential concern in any study and can result in a type of selection bias. By identifying potential differences between the target population and the study population as well as the sources of error that may occur in the data, we can better decide how to address these discrepancies to get a better estimate.[13]

2.3.4.1 Step 2: Application

The data used in our example is the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS) as used by Hernán and Robins.[4,18] NHANES-I was conducted in 1971-1975 the NHEFS was conducted in 1982-1984. The DAGs in **Figure 2.2** illustrate how the DAG evolves based on the sampling procedures and analytic decisions as well as the availability of data. For the DAGs in **Figure 2.2**, the mediators (M_1) have been greyed out since we do not have information on their diet or physical activity level post-smoking cessation our dataset. The sampling procedures for NHANES-I and NHEFS have been previously documented in greater detail elsewhere.[46,47] Complex sampling methods were utilized to properly reflect the target population. As a result, some groups were oversampled to ensure sufficient representation in the survey.[46,47] To address this, the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC) generated sampling and cluster weights to be used with the data. Both NHANES-I and NHEFS are complex survey data and as such, the sampling design should be taken into account.[48] When dealing with complex survey data, it is highly recommended to use these weights to improve generalizability, especially when addressing a descriptive aim.[46,47,49] DAGs can help determine whether using these weights will introduce bias or aid in inference.[48] **Figure 2.2a** demonstrates the potential bias being addressed by the weights to correct for oversampling, non-response, and post-stratification adjustment based on certain baseline characteristics. Due to the multistage nature of the NHANES sampling procedures, additional adjustment is needed to account for clustering and stratification.[46,47,49] As such, cluster and primary sampling unit (PSU) information is also given and should be used with the sampling weights to make inference on the survey's target population.[46,47,49] **Figure 2.2b** is a post-intervention DAG that demonstrates how using these weights together can reduce the potential bias that was introduced during the sampling procedure. By using these weights, the arrow from baseline covariates to the selection node (S=1) is removed to reflect the target population. If our target population was civilian noninstitutionalized 25–74-year-olds who smoke in the contiguous United States between 1971 and 1982 as the survey was designed, we would use these weights to more closely approximate the target population. In the current analysis, our target population is the analytic sample itself so we will not apply these weights.

The outcome of interest is weight gain at the end of the study period. Weight was measured in kilograms (kg) at baseline in the NHANES-I 1971 survey, as well as during follow-up in the NHEFS survey. Weight gain was measured by subtracting the baseline weight from the follow-up weight. As such, the change in weight variable was continuous. Percent change in weight was also assessed by dividing the change in weight by the baseline weight. In accordance with other studies that assessed weight gain, if the percent change in weight was greater than or equal to 7%, the weight gain was considered to be clinically significant.[50–52] For this example, we assess mean and standard deviations of weight at baseline, weight at follow-up, weight change, and percent weight change. We also assess frequencies of clinically significant weight gain.

Our primary covariate of interest is smoking cessation status. Smoking cessation was measured as those reporting having quit smoking between baseline and follow-up. Additionally, we are also interested in the distribution of other baseline characteristics of the entire sample as well as those who quit smoking and those who did not. Given our interest in smoking cessation and in accordance with the analysis done by Hernán and Robins, analysis was restricted to smokers whose smoking status was known at both baseline and follow-up.[4] In addition, the study population was further restricted to those having a known weight measurement, sex, age, race, height, education, alcohol use, and smoking intensity at baseline and follow-up.[4] However, by doing so, we are censoring on an event that occurred after treatment was initiated.[4,13,53–55] If the probability of lost-to-follow-up or death before the 1982 follow-up was associated with baseline covariates, smoking cessation, or potential weight gain, then selection bias could still be an issue even though the aim is descriptive because we may have induced an association between selection and the outcome.[4,13,23,53–55] The DAG in **Figure 2.2c** illustrates the potential selection bias introduced by censoring on missing data. By essentially conditioning on who participated in the 1982 follow-up survey, we may have introduced bias because we are conditioning on a potential collider. It may be beneficial to further explore this possibility in a bias analysis. It is also worth noting, that any action we take to mitigate these biases will ultimately affect how we interpret any results we get.[13] Lastly, we do not discuss it in this study but we could also have experienced measurement error due to the survey nature of the data. Although other studies should carefully consider and evaluate the presence of measurement error in any study, we will assume that any measurement error is minimal to retain simplicity. All analyses were preformed using SAS software version 9.4.[56]

2.3.5 Step 3: Specifying the measure of occurrence and role of covariates

For descriptive aims, there is a wide range of measures of occurrence to choose from. Often a researcher may only be interested in the distribution or mean of the outcome in a given population. These can be useful in assessing prevalence, incidence, or cumulative incidence in the population of interest. The nature of the data also plays a role in the selection of a measure of occurrence. For instance, whether the outcome is continuous or categorical plays a role in the selection of a measure of occurrence. If an outcome is continuous, we may choose to calculate the mean and compare it across different subsections of the population while categorical outcomes are more conducive to frequency distributions. Additionally, cross-sectional studies may be limited to assessing prevalence while longitudinal studies may be able to assess incidence rates and risk. While causal aims can struggle to assess of prevalence due to questions in temporality and the potential for reverse causation, descriptive aims are more conducive to assessing such measures because they are often designed to inform public health planning for services and interventions.[13] Although incidence and risk can also be assessed in descriptive studies, they are still subject to potential competing risks. Lastly, rates could also be of interest and are particularly useful when addressing descriptive aims as they do not necessarily require individual-level data and can be used to describe incidence over time.[13]

Specifying the measure of occurrence can lead to questions about how to deal with potential covariates. The inclusion of covariates is generally seen as adjusting on that variable to remove confounding. However, as previously mentioned, in descriptive studies confounding is less of an issue since causation is not the goal.[13,23] When answering descriptive questions, we aim to describe what we currently see in the population. We may be tempted to remove nuisance variables by adjusting on them to more clearly see the association between a covariate we are interested in and the outcome. However, this could lead to overadjustment and misinterpretation of the results as having a causal interpretation.[13,23] Additionally, instead of clarifying the association we are interested in, we may end up diluting it. Instead, we may consider stratifying or standardizing over a covariate to further investigate the association. Nevertheless, we should employ caution when deciding to adjust for any covariates with a descriptive aim to ensure that results are clear and interpreted correctly. Evaluating the DAG created in step 2 can help decipher the limitations of the data to assess a measure of interest. If we are forced to make analytic decisions due to the availability of the data, a DAG will show what avenue can and cannot be explored given the current data. It may also indicate which variables could be influenced by noise or preventing inference to the target population. In this case, standardization or stratification may be appropriate to clarify whether the association is due to its relationship with another variable or not.[13] However, any such covariates that are used in stratification or standardization should be clearly defined to aid interpretation.[13] Ultimately, however, the choice of measure of occurrence and the role of the covariates will be determined by the research question.[13,23,27]

2.3.5.1 Step 3: Application

In our example, we are primarily interested in the distribution of those who gained weight. Therefore, we are interested in proportions and means to assess the characteristics of those who gain weight compared to those who do not. In particular, we are interested in the proportion of individuals who gained weight after smoking cessation. We also are interested in the crude risk difference and mean difference of weight gain between those who quit smoking and those who did not. To assess the crude risk difference, we use a linear risk model to regress significant weight gain on smoking cessation status. Similarly, to assess the crude mean difference, we use a linear mean model to regress weight gain in kilograms on smoking cessation status. **Figure 2.2c** tells us that smoking cessation status may be influenced by baseline covariates however, they also influence weight gain, and stratifying or standardizing on these baseline covariates may result in overadjustment since we aim to describe who is gaining weight overall. Our primary issue based on this DAG is potential selection bias since we were forced to censor on missing data in the outcome. We will need to evaluate this potential selection bias in a sensitivity analysis.

2.3.6 Step 4: Estimation

In this step we estimate our measure of occurrence based on what we have learned from the previous steps. Any actions taken during the analysis should be documented in the DAG if appropriate. For instance, if adjusting or stratifying on a covariate, it should be indicated as such in the DAG to guide interpretation later on.

2.3.6.1 Step 4: Application

In our example we are assessing simple frequencies and means. We are also regressing weight gain on smoking cessation status to assess the crude risk and mean difference; however, we are not adjusting on any other variables. Thus, we do not need to make any changes to our current DAG (**Figure 2.2c**).

2.3.7 Step 5: Sensitivity/Bias analysis

Here, we use what we learned in the previous steps to assess sources of potential bias. As previously mentioned, confounding may be of less concern than addressing descriptive questions. However, descriptive aims are not immune to issues of missing data, selection bias, and measurement error.[13,23] To assess the degree of bias, we should conduct sensitivity/bias analyses to quantify them.[4,23,53,55,57] For instance, if selection is not independent of the outcome of interest or measurement of the outcome or any of the stratifying covariates has not been validated to have high sensitivity and specificity, we may wish to conduct quantitative bias analysis to assess these potential sources of bias. Similarly, missing data on the outcome or stratifying covariates can result in bias being introduced due to selection. External data may be required if the current data is not sufficient to account for the biases at play. Exploring these potential sources of bias can lend further credibility and transparency for the interpretation of our results.[4,23,53,55,57]

2.3.7.1 Step 5: Application

In our study, we can see that we may have induced selection bias by censoring on missing data through the association of censoring and the covariates (**Figure 2.2c**). As a result, the outcome is potentially not independent of selection into the study and we may need to correct for it depending on our target population. In our case, since our target population is the analytic sample itself, this would not need to be corrected. However, if we wished to generalize beyond the analytic sample to the general population then the selection bias we have induced would require correction for proper inference. In which case, to remove this selection bias, we would have to use inverse probability of censoring weighting (IPCW) to remove the bias.[55,58,54,59] To assess the potential presence of selection bias, we evaluate the distribution of smoking cessation status, sex, and weight at baseline in the censored and uncensored study population.

2.3.8 Step 6: Interpretation

In this step, we use the knowledge we gained from steps 0-5 to inform our analysis and interpretations. Covariates and outcome(s) of interest have been defined and potential sources of bias have been identified. We can use this information to inform which measure of occurrence can be estimated as well as whether we can infer anything on our target population. If selection bias or measurement error are a concern, methods traditionally used in causal studies to quantify and adjust for these biases can be used to address it.[23,30,53–55,58,60–62] The information provided to us in the DAGs generated will inform the interpretation of results and the limitations for inference.

2.3.8.1 Step 6: Application

After restricting analysis to those with complete data on covariates, the study population consisted of 1,629 smokers. An additional 63 were censored due to missing data on the outcome and were not included in analysis of weight gain. **Table 2.3** presents the unweighted frequencies and means for the study population. Due to the fact that we do not adjust for selection bias at this point, we are limited to making inferences on the analytic sample. Here, we interpret the results with respect to the analytic sample. The majority of the study population did not experience significant weight gain (n=1,201; 63.67%). Those who quit smoking comprised 26.27% of the study population (n=428). The mean age of the analytic sample was 43.9 years (95% CI = [43.3, 44.5]), while the mean weight gain was 2.6 kilograms (95% CIs: [2.2, 3.0]). When we stratified on significant weight gain (i.e., an increase in weight at follow-up of 7% of the baseline weight), those who experienced significant weight gain were majority female (n=312; 54.83%), younger (mean=40.9; 95% CI = [40.0, 41.8]), very active on a daily basis at baseline (n=273; 47.98%), and had a lower mean weight at baseline (67.4 kg; 95% CI = [66.3, 68.6]). Additionally, the mean years of smoking were lower among those that had significant weight gain than those that did not have significant weight gain (22.3 years versus 25.9 years, respectively). When assessing smoking cessation status, a higher proportion of those who had significant weight gain had quit smoking compared to those who did not have significant weight gain (30.76% versus 22.87%, respectively).

Table 2.4 displays the results of the linear risk and mean regression to assess the crude association of smoking cessation on weight gain. Those in the study population who quit smoking had a lower risk of significant weight gain than those who did not quit smoking. However, those who quit smoking experienced an average weight gain that was 2.54 kg higher than those who did not quit smoking.

To assess the presence of selection bias induced by censoring on those who had unmeasured weight at follow-up, we first checked the distribution of smoking cessation status, sex, and baseline weight among smokers in the censored versus uncensored study population (**Table 2.5**). There are differences in the distribution of smoking cessation status and sex between the uncensored and censored population. When we look at weight at baseline, it appears that on average the censored population is heavier than the uncensored population however, the overlap of the confidence intervals indicates that the means may not be that different. Nonetheless, there is potential for selection bias to be an issue and should be accounted for if the target population is not the analytic sample itself.

2.4 Conclusion

Descriptive aims play a key role in epidemiology for hypothesis-generation, identifying health disparities, and targeting interventions by characterizing what we see in the population of interest. Due to the fact that descriptive aims heavily rely on clear definition of the target population, the outcome, and the role of covariates, it is especially important to understand how data collection, measurement, and analysis affect our ability make inference as we would when addressing causal aims. We have walked through the steps and considerations on how to address descriptive aims and how to use DAGs to direct the analysis and interpretations. The limitations of DAGs are well-known as DAGs rely heavily on the assumption of faithfulness and may not be able to fully illustrate parametric concepts such as effect modification.[2,63] However, DAGs are still flexible tools that can be augmented and evolve to reflect these concepts. Therefore, DAGs remain useful tools to deal with missing data, assess sources of potential bias, identify critical variables for analysis and/or data collection, communicate assumptions, and guide interpretation even when the aim is descriptive.

2.5 Appendix

Box 2.1: Description of person, place, and time

Person - Who is affected? What defining characteristics do they have in common? Who is *not* affected? How many are affected?

Place - What areas have higher disease frequency? What is unique about those areas?

Time - How often does exposure/disease occur? How does exposure/disease frequency change over time? What else could be associated with those changes?

Box 2.2: Four purposes of descriptive aims driven by the person, place, and time

1) Generate new hypotheses for analytic studies
2) Identify and assess health disparities
3) Identify potential areas for etiologic studies and subsequent targeted intervention
4) Describe disease or exposure burden and assess health needs

Table 2.1: Checklist items developed by Lesko et al.[13] for reporting of descriptive epidemiologic studies

Section	***Recommendation***	***Use of DAGs***
Title and Abstract	1. Explicitly state the study goal is description in the title or the abstract.	No, not required. A basic DAG consisting of the outcome and a primary covariate (if applicable) of interest can be drawn but is not necessary.
	2. Summarize the target population and provide an informative and balanced summary of estimated disease occurrence in the abstract.	No, not required. A basic DAG consisting of the outcome and a primary covariate (if applicable) of interest can be drawn but is not necessary.
Background./rationale	3. State the motivation for the study including, where relevant, the action that might be informed by the results.	Yes, display available background knowledge. Encode relationships between variables in the DAG.
Objectives	4. State the descriptive estimand, explicitly including: a) the target population b) the health state to be summarized c) the measure of occurrence d) any stratification variables, if applicable	Yes, use the DAG to identify stratification variables, define exposure/outcome variable, and assess the appropriateness of the measure of occurrence
Study design	5. Study design: a) State whether the study is cross-sectional or longitudinal b) Restate the measure of occurrence being targeted. c) If the study is longitudinal, specify the time origin and follow-up period for the measure of occurrence; if the study is cross-sectional, specify the time-anchor at which the health state is summarized for individuals	Yes, use the DAG to display time-points and assess the appropriateness of the measure of occurrence
Setting	6. Describe any relevant features of the place and time in which data were collected	Yes, display background knowledge. Encode relationships between variables in the DAG.
Participants	7. Participants: a) Describe the target population thoroughly in terms of person, place, and time b) Describe sampling into the study population (whether sampling was explicit or implicit, e.g., by inclusion in an administrative database); this includes eligibility criteria. c) Describe any restrictions on the analytic sample	Yes, display background knowledge and include inclusion/exclusion criteria for the study population. Assess the presence of selection bias.

Section	***Recommendation***	***Use of DAGs***
Outcome(s)	8. Outcome(s): a) State when and how the outcome is measured b) Include estimates or discussion of sensitivity and specificity of the study outcome definition relative to the gold standard c) List secondary outcomes or competing events of interest	Yes, use the DAG to display sources of missing data and measurement error/information bias. Identify competing events.
Covariates	9. Specify any stratification or adjustment variables -- clearly define how variables were collected or constructed	Yes, use the DAG to identify any stratification or adjustment variables
Data sources/measurement	10. Clearly delineate any inclusion/exclusion criteria for membership in the data source, including the original purpose for which the data were collected, if not for the study at hand	Yes, display inclusion/exclusion criteria for the study population. Assess the presence of selection bias.
Bias	11. Describe any assumptions or methods used to extrapolate data from the analytic sample to the study population, and from the target population	Yes, use the DAG to display assumptions and assess the presence of selection bias, information bias, and/or missing data issues.
Statistical methods	12. Statistical methods: a) Describe the primary statistical methods used to estimate the measure of disease occurrence being targeted; discuss assumptions of that method in light of data limitations (e.g., assumption of independent censoring for people lost to follow-up) b) If any adjustment/standardization will be done, state the goal of such adjustment	Yes, display inclusion/exclusion criteria for the study population. Use the DAG to identify any stratification or adjustment variables. Augment the DAG to include consequences of analytic decisions
Participants	13. Report numbers of individuals at each study stage	Yes, display inclusion/exclusion criteria for the study population.
Descriptive data	14. Descriptive data: a) Report on the characteristics of the analytic sample in a "table 1" b) Indicate the number of participants with missing data for each variable used in the analysis c) If any weighting or imputation is done to reconstruct the study sample or target populations, include columns for those populations	Yes, augment the DAG to include any analytic decisions that were made for interpretation later
Outcome data	15. Outcome data: a) Present an overall (unstratified estimate of the measure of occurrence of interest b) Report "crude" (raw data in the analytic sample) and (if applicable) "corrected" (after any weighting or imputation)	Yes, augment the DAG to include any analytic decisions that were made for interpretation later

Section	***Recommendation***	***Use of DAGs***
Other analyses	16. Present pre-specified stratum-specific or adjusted/standardized results	Yes, augment the DAG to include any analytic decisions that were made for interpretation later
Key results	17. Summarize key results with reference to the study objective	Yes, use the DAG to guide the interpretation based on whether the target population can be inferred on.
Limitations	18. Summarize potential sources of selection bias and/or measurement error and any attempts to mitigate these biases; Discuss both direction and magnitude of any potential bias; Integrating quantitative bias analysis in the study is encouraged	Yes, use the DAG to guide the interpretation based on whether the target population can be inferred on.
Interpretation	19. Interpretation: a) Avoid causal interpretation of descriptive results; Avoid over-interpreting stratum-specific difference in measures of occurrence b) Describe how results of this study might inform or improve public health or clinical practice	Yes, use the DAG to guide the interpretation based on whether the target population can be inferred on.

Table 2.2: Outline of steps for addressing a descriptive aim and the use of Directed Acyclic Graphs (DAGs)

Step	*Description*	*Lesko et al. Checklist Item(s)*[13]	*Use of DAGs*
0. Define the research question	Define the research question. Define the target population. Define the rationale for the aim (i.e. what is the objective of the aim?). Specify the measure of occurrence. Define the outcome(s). Specify the time period.	1-4, 7a	Not necessarily applicable - this step is preparation for DAG creation and scope of study. A basic DAG consisting of the outcome and a primary covariate (if applicable) of interest can be drawn but is not necessary.
1. Specify knowledge about the system	Describe available background knowledge with respect to the question under study	2, 3, 4, 6, 7a	Display background knowledge. Encode relationships between variables in the DAG.
2. Link the data to the system	Identify which variables have been measured. Describe selection/inclusion criteria. Link the observable data to the system. Define the sample/study population. Evaluate sensitivity and specificity of the exposure/outcome measurement.	5a, 5c, 7b, 7c, 8-11	Identify sources of selection bias and/or measurement error. Assess issues with missing data. Augment the DAG to specify what data is observable vs unobservable and where error may exist. If data is still needed to be collected, use the DAG created in the previous step to select most high yielding variables and include limitations of the data collection procedures.
3. Specification of the target measure of occurrence & the role of covariates	Identify the measure of occurrence of interest. Specify the role of covariates.	5b, 9, 12	Use the DAG to identify covariates that could determine the distribution of the outcome. Assess whether the data is compatible with the measure of occurrence.
4. Estimate	Estimate the measure of occurrence in the analytic sample.	13-16	Augment the DAG to incorporate and analytic decisions made for inference later
5. Sensitivity/Bias analysis	Test sensitivity of the measure of occurrence due to assumptions made or potential bias	11, 14c, 15b, 18	Guide the sensitivity or bias analysis using a DAG to assess sources of selection bias or measurement error/information bias

Step	***Description***	***Lesko et al. Checklist Item(s)***[13]	***Use of DAGs***
6. Interpretation	Assess what assumptions are required to infer from the analytic sample to the study population and the target population. Interpret the estimate in light of the prior steps.	17-19	Use the DAG to guide the interpretation based on whether the target population can be inferred on. Clarify the assumptions used to apply the data from the analytic sample to the study population and target population

Box 2.3: Four characteristics of a well-defined research question for descriptive aims

1) *Target population* - a population that is grounded in time and specifies the person and place of interest

2) *Exposure or Outcome(s)* - a defined event, disease, health state, or characteristic of interest

3) *Measure of occurrence* - a target measure of interest that aims to quantify or summarize the distribution of the outcome in the target population

4) *Potential covariates* - any variables that may need to be stratified or standardized on to mimic the distribution of the target population

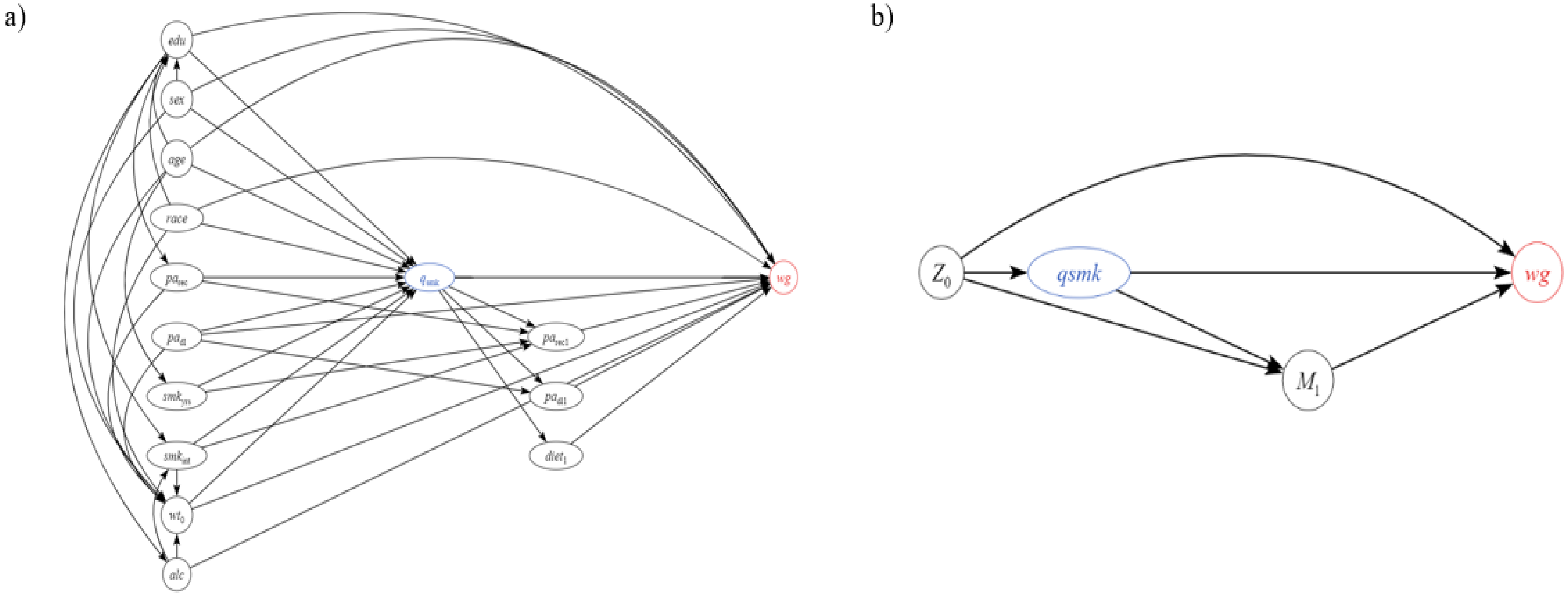

Figure 2.1: Potential directed acyclic graphs (DAGs) for the relationships between baseline covariates and the outcome(s) of interest. a) This DAG shows the relationships between all potential covariates and weight gain (*wg*). Baseline covariates include: education (*edu*), sex (*sex*), age (*age*), race (*race*), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), smoking intensity (smk_{int}), years of smoking (smk_{yrs}), weight at baseline (wt_0), and alcohol frequency (*alc*). The main covariate of interest is smoking cessation status (q_{smk}). The potential mediators are post-smoking cessation recreational physical activity (pa_{rec1}), post-smoking cessation physical activity in daily life (pa_{dl1}), and post-smoking cessation diet ($diet_1$). b) This DAG also shows the relationships between baseline covariates (Z_0), smoking cessation, the post-smoking cessation mediators (M_1), and weight gain. However, this DAG also assumes we do not care about the relationships between the baseline covariates.

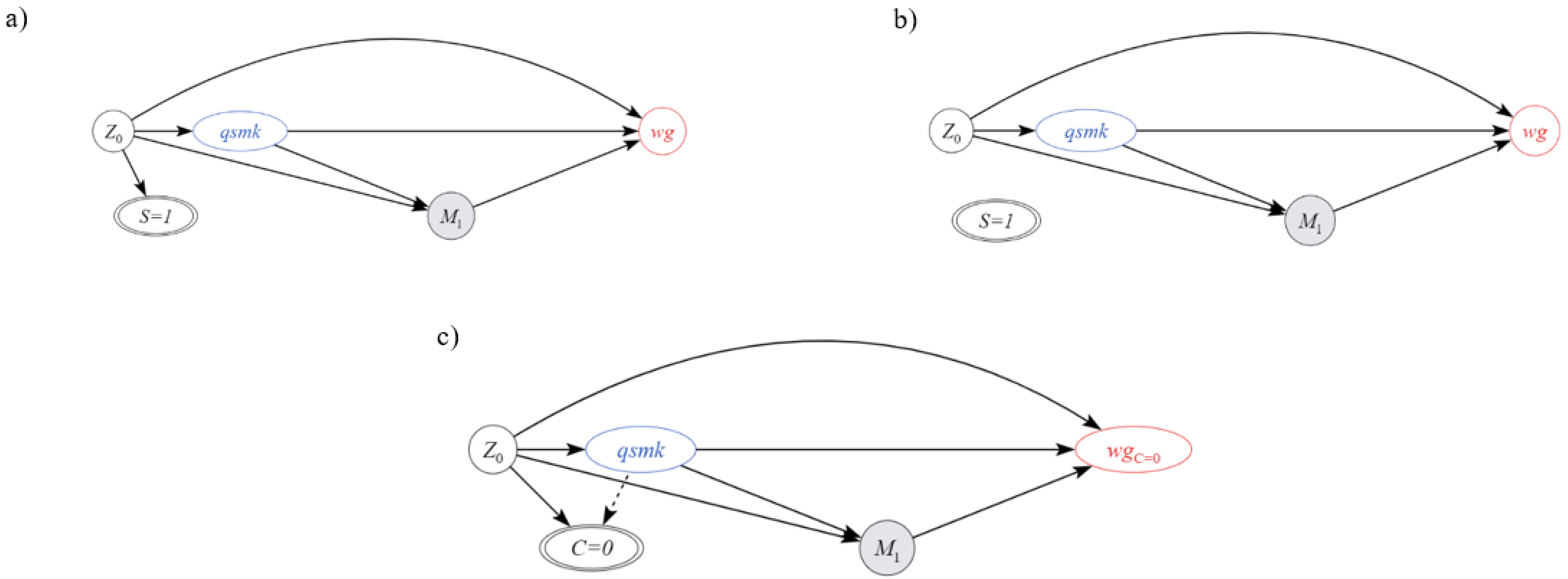

Figure 2.2: Potential directed acyclic graphs (DAGs) for the relationships between covariates and the outcome(s) of interest in the context of the data. a) This DAG shows the relationships between the baseline covariates (Z_0), smoking cessation (q_{smk}), the unmeasured post-smoking cessation mediators (M_1), and weight gain (*wg*). Baseline covariates include: education, sex, age, race, recreational physical activity, physical activity in daily life, smoking intensity, years of smoking, weight at baseline, and alcohol frequency. This DAG also includes a selection node (*S=1*) to represent the sampling procedures. b) This is an intervention DAG to depict how sampling weights affect the DAG in a) and improves inference to the target population. c) This DAG is similar to a) but depicts selection bias induced by censoring on measured weight in the 1982 follow-up. The dotted line represents the potential association between smoking cessation and censoring.

Table 2.3: Frequency and means of characteristics for those with significant weight gain and. those without significant weight gain, NHEFS 1982-1984

Characteristic	*Total sample population (N=1,629)*	*Significant weight gain (N=569)*	*Not significant weight gain (N=997)*
	n (%)	*n (%)*	*n (%)*
Significant weight gain			
Yes (≥7%)	569 (36.33%)	569 (100.00%)	--
No (<7%)	997 (63.67%)	--	997 (100.00%)
Quit smoking			
Yes	428 (26.27%)	175 (30.76%)	228 (22.87%)
No	1,201 (73.73%)	394 (69.24%)	769 (77.13%)
Sex			
Male	799 (49.05%)	257 (45.17%)	505 (50.65%)
Female	830 (50.95%)	312 (54.83%)	492 (49.35%)
Race			
White	1,414 (86.80%)	497 (87.35%)	863 (86.56%)
Black or other	215 (13.2%)	72 (12.65%)	134 (13.44%)
Education			
8th Grade or less	311 (19.09%)	81 (14.24%)	210 (21.06%)
High School Dropout	351 (21.55%)	135 (23.73%)	205 (20.56%)
High School	659 (40.45%)	240 (42.18%)	397 (39.82%)
College Dropout	126 (7.73%)	52 (9.14%)	69 (6.92%)
College or more	182 (11.17%)	61 (10.72%)	116 (11.63%)
Exercise at baseline			
Much exercise	317 (19.46%)	115 (20.21%)	185 (18.56%)
Moderate exercise	677 (41.56%)	241 (42.36%)	420 (42.13%)
Little or no exercise	635 (38.98%)	213 (37.43%)	392 (39.32%)
Daily activity at baseline			
Very active	729 (44.75%)	273 (47.98%)	429 (43.03%)
Moderately active	738 (45.30%)	237 (41.65%)	478 (47.94%)
Inactive	162 (9.94%)	59 (10.37%)	90 (9.03)
Alcohol frequency at baseline			
Almost every day	336 (20.69%)	111 (19.61%)	214 (21.51%)
2-3 times/week	231 (14.22%)	75 (13.25%)	144 (14.47%)
1-4 times/month	506 (31.16%)	189 (33.39%)	305 (30.65%)
< 12 times/year	344 (21.18%)	134 (23.67%)	194 (19.50%)
No alcohol last year	207 (12.75%)	57 (10.07%)	138 (13.87%)

Characteristic	***Total sample population (N=1,629)***	***Significant weight gain (N=569)***	***Not significant weight gain (N=997)***
	Mean (95% CI)	*Mean (95% CI)*	*Mean (95% CI)*
Age (years)	43.9 (43.3, 44.5)	40.9 (40.0, 41.8)	45.2 (44.5, 46.0)
Weight at baseline (kg)	71.1 (70.3, 71.8)	67.4 (66.3, 68.6)	72.8 (71.8, 73.8)
Weight at follow-up (kg)	73.5 (72.7, 74.3)	77.4 (76.0, 78.7)	71.2 (70.3, 72.2)
Change in weight (kg)	2.6 (2.2, 3.0)	10.0 (9.5, 10.4)	-1.5 (-1.9, -1.2)
Percent change in weight (%)	4.2 (3.7, 4.8)	15.0 (14.3, 15.7)	-1.9 (-2.4, -1.5)
Smoking intensity at baseline (cigarettes/day)	20.6 (20.0, 21.1)	20.3 (19.3, 21.2)	20.7 (19.9, 21.4)
Change in smoking intensity (cigarettes/day)	-4.7 (-5.4, -4.1)	-5.5 (-6.7, -4.4)	-4.1 (-5.0, -3.3)
Years of smoking	24.9 (24.3, 25.5)	22.3 (21.4, 23.2)	25.9 (25.1, 26.7)

Table 2.4: Weighted and unweighted crude risk and mean differences of weight gain on smoking cession status, NHEFS 1982-1984.

	Significant weight gain (linear risk model) *Estimate (95% CI)*	***Change in weight (linear mean model)*** *Estimate (95% CI)*
Quit smoking	-0.10 (-0.15, -0.04)	2.54 (1.66, 3.42)

Table 2.5: Checking for selection bias by assessing frequency of smoking cessation status in censored versus uncensored population, NHEFS 1982-1984.

Characteristics		***Uncensored Population (C=0)***	***Censored Population (C=1)***
Quit Smoking: *n (%)*			
	Yes	403 (25.73%)	25 (39.68%)
	No	1,163 (74.27%)	38 (60.32%)
Sex: *n (%)*			
	Male	762 (48.66%)	37 (58.73%)
	Female	804 (51.34%)	26 (41.27%)
Weight at baseline (kg): *Mean (95% CI)*		70.83 (70.07, 71.59)	76.55 (70.67, 82.43)
TOTAL: *n (%)*		1,566 (100.00%)	63 (100.00%)

Chapter 3 DAGging out predictive aims

3.1 Abstract

Prediction is often used in clinical and public health settings to predict disease diagnosis, prognosis, or potential outcomes. Prediction is largely viewed as separate from causal inference and indeed some forms of prediction require more understanding of the causal structure than others. Thus, the use of directed acyclic graphs (DAGs) in the development of these types of models is rare. Here, we adapt previous frameworks for the development of predictive models to provide guidance on the use of DAGs in the development of predictive models. We also demonstrate one application of this adapted framework that uses prediction to rank the variable importance for the prediction of weight gain in the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS). This chapter is intended to provide guidance on the use of DAGs in predictive studies to optimize variable selection, address sources of bias and missing data, as well as communicate key characteristics that may facilitate the transportability of the model to other populations.

3.2 Introduction

In the age of the internet, the amount of data on individuals has grown exponentially. As computers are able to share information more rapidly and efficiently, data on nearly every aspect of a person’s life is collected. Researchers have aspired to use this data to predict and describe behavior and outcomes for a variety of reasons such as movie recommendations, currently trending topics, economic markets, and business performance. Health researchers and epidemiologists may also use big data for prediction purposes.[22,30] There are generally two forms prediction can take in epidemiology. The first is counterfactual prediction which requires causal inference as it aims to predict a potential outcome based on some intervention.[22,64] Counterfactual prediction aims to answer the “what if we had done” questions that researchers may have to assess the outcome of an intervention had it taken place.[22,64] In Pearl’s ladder of causation, these questions fall on the third rung of counterfactuals as they require us to *imagine* what may have occurred had some intervention taken place counter to fact.[19] The second form of prediction in epidemiology falls under the umbrella of clinical prediction modeling. Questions under this umbrella often aim to predict events.[14,30] These questions are either diagnostic or prognostic in nature. Where a diagnostic question aims to predict whether an individual is likely to *have* a certain disease/condition based on current characteristics, a prognostic question aims to predict whether an individual is likely to *acquire* a certain disease/condition in the future based on current characteristics.[14,30] Prognostic predictive models generally aim to prevent the disease. Prognostic models may still require some form of causal inference as various treatment options are considered.[20] Diagnostic predictive models generally remain on the first rung of the ladder of causation since we are predicting the outcome of interest based on what we *see* while prognostic predictive models can move into the second rung since we are predicting the outcome based on what is *done* now and counterfactual prediction is solidly on the third rung since it requires us to re-*imagine* the past.[19]

As described in Chapter 1, clinical predictive models generally aim to increase the *d*-connectedness between the treatment/exposure of interest. Thus, understanding and establishing the relationships between the outcome and any potential predictors can be useful in order to minimize the number of predictors in the model to ease interpretability and increase the transportability of the model. Piccininni et al. has previously described the usefulness of directed acyclic graphs (DAGs) for the development of diagnostic predictive models. They demonstrate the efficiency of the use of the Markov blanket to select predictors for the model to improve calibration, increase transportability, and ease interpretation through simulation. Here, we demonstrate how to incorporate the use of DAGs throughout the development of a predictive model using and expanding on the steps for development outlined by Steyerberg and Vergouwe with application to empirical data taken from *Causal Inference: What If*.[4,14,15,18]

3.3 Developing a clinical predictive model using DAGs

3.3.1 Overview of the steps for predictive model development

To provide a logical framework to improve the quality of clinical predictive models, Steyerberg and Vergouwe outlined seven steps for the development of such models (**Table 3.1**).[14,15] These steps include: 1) specifying the research question; 2) defining the coding of predictors; 3) model specification; 4) model estimation; 5) model performance; 6) model validation; and 7) model presentation.[14,15] We adapt these steps to incorporate the use of DAGs in predictive model development. **Table 3.2** provides an overview of how these steps fit into our proposed framework and includes additional steps that may be useful to increase transparency and reporting accuracy using DAGs.

3.3.2 Step 0: Specify the research question

As with any study, the most important step to start with is to identify the research question. Part of doing so, as previously mentioned, is to establish whether the question has a predictive aim. If so, it is important to identify whether the aim is prognostic or diagnostic. Is the interest in predicting the future outcome based on a certain course of action given current conditions (prognostic) or is it in predicting the likelihood of disease state given current conditions (diagnostic)? Similarly, does the question have clinical relevance? Answering these questions requires the researcher to define the target population, specify the intended use of the model, and clearly define the outcome (including the endpoint). Additionally, if the question is prognostic, defining the treatment of interest may be required to identify an appropriate approach to properly support treatment decisions.[20]

3.3.2.1 Step 0: Application

To demonstrate the application of the use of DAGs in creating a prediction model, we will aim to identify the greatest predictors of weight gain in the study population. Often in public health, we wish to identify points of potential intervention or means of diagnosis. To do so we need to understand what contributes to the outcome. Doing so may also influence future data collection efforts and points of study. In our example, our target population are smokers in the study population. Our outcome is weight gain at follow-up which is approximately 11 years after baseline. The main aim is to rank which predictors are important to predict future weight gain.

3.3.3 Step 1: Specify knowledge about the system

Often in predictive studies, there is limited knowledge available. Additionally, the causal structure defining the relationship between the predictors and the outcome may not traditionally be of huge concern. However, when strong knowledge of the structure of the system is available, it is beneficial to present the a priori assumptions that may influence predictor selection in a DAG. An initial DAG based on the available knowledge should be generated based on a review of the literature and expert input to display background knowledge. The DAG should provide insight on what is already known about the predictors, if there are any interactions or sources of heterogeneity we might be concerned with. Guidance on the identification and incorporation of potential sources of heterogeneity in a DAG has be described elsewhere.[65–67] Often the goal of prediction is the development of a model that is generalizable and transportable to other populations. Thus, it is important to know what predictors could be of value beyond what is available in the data. If intuitive assumptions are made in addition to the assumptions that are knowledge-based, they should be clearly depicted in the DAG as distinct from those that are knowledge-based. Alternatively, a purely knowledge-based DAG should be created in addition to the DAG that includes intuitive assumptions. Different predictive models require different levels of causal structure to properly address the question.[20,22] For instance, in some prognostic models and counterfactual prediction models, a causal structure is necessary to answer the research question whereas in diagnostic models, correlation is all that is desired but a causal structure could optimize model specification and improve validity.[20,22,32]

3.3.3.1 Step 1: Application

The DAG depicted in **Figure 3.1** shows the potential relationships between the baseline predictors and the outcome of interest. The DAG shows us that the predictors in the graph are more likely to be associated with the outcome of interest, especially those with a direct connection to the outcome or colliders. Since our aim is to establish which predictors have high importance so that we may potentially explore intervention on these predictors to prevent future weight gain, we would also like to ensure that temporality holds. The DAG helps us isolate which predictors should be considered when developing our model. In our DAG, we can already see that smoking cessation (q_{smk}), education (*edu*), age, smoking intensity (cigarette/day) at baseline (smk_{int}), and weight in kilograms (kg) at baseline (wt_0) will be critical to our model as they lie on more paths than other nodes in the DAG. Similarly, post-smoking cessation recreational and daily life physical activity as well as diet post-smoking cessation are also valuable since they have the most direct connection to weight gain in kg at follow-up temporally.

3.3.4 Step 2: Link the data to the system

Before we can jump into specifying our model for prediction, we first need to understand the data we are working with. Part of understanding the data includes knowing the underlying study design, defining the inclusion/exclusion criteria used to select observations, and identifying which potential predictors have been measured. By defining the inclusion/exclusion criteria, we can clarify the sample/study population. We can do an exploratory analysis to assess whether we have missing data issues and better understand and prepare the data for analysis by assessing the coding of the variables. Specifically, when assessing the coding of predictors, there are various ways in which continuous and categorical variables can be coded. We may determine that some variables require recoding for more meaningful interpretation or collapsed due to infrequency in some categories.[14,15] Additionally, some variables may require transformation or to account for heterogeneity. If this is the case, it is important to include these terms in the DAG for interpretation later. For missing data, we can assess the nature of missingness to determine what method is best suited for treating that missing data. Simple imputation is often used in packages to fix missing data issues in standard statistical packages but may still result in a biased estimate and reduce external validity of the estimate if the causal structure of the missingness is not considered.[68,69] Proper documentation of missing data can also aid other researchers to assess the validity of the model in the future for use in other population. To properly document the nature of the missing data, we would assess missingness patterns and depict the causal structure of the missing data.[69–71] To depict the causal structure of missingness in DAGs, previously published guidance is available.[29,72,73] Additionally, depending on the assumptions required to address the research question, confounding or confounding by indication can be an issue that should be assessed.[9,20] Selection bias and measurement error are of concern with prediction especially when missing data is an issue. DAGs will be useful to detect the potential presence of these biases so that they can be properly addressed.

3.3.4.1 Step 2: Application

For our example, we used the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS) as provided by Hernán and Robins.[4,17,18,46,47] The data provided is restricted to answered the medical history questionnaire at baseline, and those with known age, sex, race, weight, height, education, alcohol use smoking intensity at baseline and follow-up. Baseline surveys were conducted in 1971 to 1975 and follow-up survey were administered to surviving baseline participants or their relatives in 1982 through 1984.[17] The sampling procedures for the survey have been described in detail elsewhere.[46,47] The original data includes sampling weights and cluster weights to account for oversampling and may need to be used if we are trying to make inference on the survey's target population. For our purposes, we will not be using these weights since our target population is not that of the survey but strictly limited to those in the analytic sample for simplicity. Serious consideration should be given about whether to use the weights or not depending on the study's target population.

The data were restricted to 1,629 smokers and when exploring the data, we were forced to remove an additional 63 smokers due to missing data in follow-up weight. The DAG in **Figure 3.2** links the data to our DAG in step one. The censoring of missing data in the outcome likely and restriction to the complete data in baseline predictors may limit our ability to generalize and transport our results to other populations. Additionally, post-cessation predictors are unmeasured and cannot be included in our analysis. As a result, smoking cessation status is now more critical as a predictor according to our DAG to predict weight gain at follow-up. Risk factors for weight gain at follow-up also include baseline predictors: alcohol frequency, weight, smoking intensity, years of smoking, recreational physical activity, daily life physical activity, age, sex, race, and education. For simplicity, we assume no measurement error in the data. However, given the nature of survey data with respect to questions on health behavior. This assumption may not be realistic and should be explored thoroughly in any prediction study. Lastly, since we are interested in identifying important predictors, we less concerned about confounding.

We ran an exploratory analysis of the data to assess potential issues with coding and check predictor balance (**Table 3.3**). We assessed the distribution of characteristics in the study population to check for balance and missing values. There were no missing data found in our predictor set. We randomly split the data into a training set and a test set where the training set contained 70% of the data and the test set contained the remaining 30%. Categorical variables were converted into dummy variables and all predictors were normalized to have values between

0 and 1 in both data sets. All analyses were performed using R Statistical Software (v4.3.0; R Core Team 2023) and the caret package was used to prepare and model the data.[74,75]

3.3.5 Step 3: Specification of the target quantity, model specification, and model estimation

Here, we respecify whether predictors or predicted individual risk is of interest. We also select predictors to be included in the model taking into consideration what assumptions need to be met for the type of model we selected. Often in prediction models, we would like to optimize the number of predictors included in the model to improve performance and increase interpretability.[32] Additionally, overfitting can reduce generalizability of the model. To do so, stepwise selection and shrinkage methods are often used to select predictors for inclusion in the model and mitigate overfitting.[14,15] However, the over-reliance of predictor selection methods should be avoided in favor of consideration of prior knowledge.[15] In the case of diagnostic prediction or questions that aim to identify and quantify important predictors, the Markov Blanket can be used to optimize variable selection.[32] Briefly, the Markov Blanket includes all parents of the outcome node, all of its children, and parents of its children.

In the case of counterfactual and prognostic prediction models, assessment of the conditions of identifiability may be required to ensure appropriate predictions. Other assumptions include identifying sources of heterogeneity if regression models will be used. To soften the additivity/heterogeneity assumption, we can include an interaction term in the model if we know the source of heterogeneity.[14,15] We again refer to the DAGs created in steps 1 and 2 to assess sources of heterogeneity so that interaction terms can be included in the model.[65–67]

3.3.5.1 Step 3: Application

With prediction models, the aim for predictor selection is to enhance the connection between the predictors and the outcome. Referring back to the DAGs we created in steps 1 and 2, we expect that smoking cessation and weight at baseline will have high connectivity with the outcome in our data since parents of weight gain at follow-up and technically colliders in the graph. Additionally, education, age, sex, race, smoking years, and alcohol frequency are also parents of the outcome so they should also have some connectivity in our data. Including recreational and daily life physical activity at baseline may be critical as they are parents of unmeasured post-smoking cessation predictors that are in turn parents of the outcome. In fact, we expect smoking cessation to have high connectivity to the outcome as well since it is the child of several other baseline predictors and a parent of the outcome. The only thing our DAG does not tell us is the magnitude of the connections.

Machine learning and predictive models have built-in mechanisms that are used to evaluate large data and identify the most important predictors.[76] For the purposes of demonstration, we will use a conditional inference random forest model to rank the importance of predictors. Random forests are non-parametric models. They are often used to assess variable importance because they are more robust than decision trees and more flexible than traditional regression models. Conditional inference random forests are generally better than regular random forests for variable importance since random forests tend to be biased towards predictors with more categories.[77] Shrinkage methods were not used. Instead, we increase the tune length to 10 to tell the algorithm to try 10 values for the number of variables that are randomly sampled as candidates at each split (mtry). The accuracy of our model is limited in part due to the nature of the data and the choice to use a random forest model. Part of the issue is that our outcome is continuous and to decide the direction at each spilt, the continuous outcome is categorized and can lose information. We will check the accuracy of our model in the next step.

3.3.6 Step 4: Model performance & validation

Model performance and model validation concern the overall fit of the model, as well as its internal and external validity. To assess model performance, several measures are traditionally used to compare model fit across different types of models.[14,15] Prediction accuracy is also considered here by comparing predicted values to observed values. This includes assessing calibration, discrimination, and clinical usefulness.[14,15] To improve internal validity, cross-validation and bootstrap resampling should be attempted. If selection bias, measurement error/misclassification bias, or confounding are an issue, consider quantitative bias methods to assess the magnitude of the bias using regression methods.[9,53,54,57,60,78] The generalizability and transportability of a predictive model is a stronger test.[14,15] The DAG created in step 2 should be considered to assess the limitations of the data to make inference on the target population. Understanding the underlying causal structure can help assess transportability. If we are attempting to transport the model to a different target population, a DAG relevant to the new population should be created and compared to assess whether transportability is possible.[33] If a critical predictor is not possible in the new population, transportability may not be possible. However, if the differences between the original target population and the new target population can be quantified, then transportability may be possible through data fusion or through the use of prediction error modifiers.[33–35] Use of these methods require identifiability conditions to be met.[33] Specifically, the positivity condition and the independence of the outcome and the target population given a set of covariates.[33–35] The use of DAGs to assess and address the issue of transportability have been previously discussed in detail.[33] In cases where the predictor is an effect of the outcome, transportability will suffer unless the predictor and outcome share at least one common cause.[32] Predictor selection with consideration to the underlying causal structure may improve the transportability of the model to other populations.[32]

3.3.6.1 Step 4: Application

We performed repeated 5-fold cross validation with 5 repetitions to train and tune the model using the training data. The root mean square error (RMSE) was compared across model iterations to select the best model (**Figure 3.3a**) and a calibration plot was generated to compare the predicted values of weight gain to the observed values by applying the final model to the test data (**Figure 3.3b**). Although our model may be internally valid, when we assess the DAG in **Figure 3.2**, we recognize that our model in only valid for the study population due to censoring. If we were to use the results of the random forest to implement a regression model, we would need to consider correcting the potential selection bias using quantitative bias analysis methods. This could involve the use of external data or further exploration into the nature of the missingness in the outcome. Additionally, when we compare the RMSE for the training data to the RMSE for the testing data. The RMSE for the final model which randomly samples 10 variables as candidates at each split is 7.12 in the training set and 8.21 in the testing set. The mean predicted weight gain is 2.6 in the training set and 2.5 in the testing set. The calibration plot reveals that the predicted values for the testing data are in pretty close agreement with the observed values.

3.3.7 Step 5: Interpretation of results

Here, we present the results of our efforts in steps 0-4. Models should be presented in a manner that is appropriate to the target audience.[15] This may include the model formula, charts, dashboards, or applications.[15] Additionally, the target population should be reiterated and inferential uncertainties and limitations should be clearly communicated to mitigate misinterpretations in the future.

3.3.7.1 Step 5: Application

Figure 3.4 displays the results of the variable importance analysis from the cross-validated conditional inference random forest. The top 5 variables of importance were, age, weight at baseline, quitting smoking, not quitting smoking, and number of years smoking. Overall, age contributed the most value for prediction of weight gain. The importance of smoking cessation status may require more causal exploration. Due to censoring of missing data in the outcome, we may have induced selection bias and thus, these results may not be generalizable beyond the analytic sample. Our DAG is consistent with these results.

3.4 Conclusion

Prediction aims are often used in clinical epidemiology for diagnosis, prognosis, and intervention. Even in pure prediction settings where associational relationships are the goal, causal considerations may still be required to improve the generalizability and transportability of the model. Depending on the predictive aim, they are not immune to concerns with selection bias, measurement error/misclassification bias, confounding, and missing data. DAGs provide a framework to assess these causal considerations and may also provide a means to optimize variable selection when dealing with big data.[20,32] We provide an example of how to incorporate DAGs in model development. Further theoretical development of the use of DAGs to address specific types of predictive aims (particularly prognostic aims) is warranted.

3.5 Appendix

Table 3.1: Steyerberg and Vergouwe's seven steps for development of clinical prediction models and the use of directed acyclic graphs (DAGs)

Step	***Considerations***	***Use of DAGs***
1. *Problem definition and data inspection*	Is the aim to provide insight on predictors or predict risk? Consider selection, predictor definitions, completeness, and endpoint definition	Yes, define nodes and display background knowledge. Encode relationships between variables in the DAG. Use a DAG to display the selection criteria and the relationship between the sample and target population. Display inclusion/exclusion criteria. Display potential sources of measurement error/information bias and missing data.
2. *Coding of predictors*	Are the predictors continuous or categorical? Should categorical variables be collapsed? Are transformations needed?	Yes, augment to DAG to include any transformations/interactions that need to be included in the model to prevent violation of model assumptions.
3. *Model specification*	What are the main effects? Do the assumptions hold?	Yes, use the DAG to optimize predictor selection.
4. *Model estimation*	Are statistical shrinkage methods required to limit overfitting the model?	Yes, use the DAG to optimize predictor selection and potentially inform shrinkage methods, if appropriate.
5. *Model performance*	Were calibration, discrimination, and/or clinical usefulness methods used?	Yes, use the DAG to address sources of measurement error/information bias that can affect sensitivity and specificity.
6. *Model validation*	Is the model internally valid?	Yes, use the DAG to assess sources of bias.
7. *Model presentation*	Is the model presented in a format appropriate for the audience?	Yes, use the DAG to guide the interpretation.

Table 3.2: Outline of predictive modeling steps and the use of Directed Acyclic Graphs (DAGs)

Step	*Description*	*Use of DAGs*
0. Define the research question	Define the research question. Define the target population.	Define potential nodes. DAGs can depict the data-generating process for the study.
1. Specify knowledge about the system*	Describe background knowledge with respect to the question under study	Display background knowledge. Encode relationships between variables in the DAG.
2. Link the observed data to the system*	Link the observable data to the DAG. Identify which predictors have been measured. Describe selection/inclusion criteria. Define the sample/study population. Do an initial inspection of the data. Define coding of predictors.	Use a DAG to display the selection criteria and the relationship between the sample and target population. Identify measured and unmeasured predictors. Explore potential sources of missing data, confounding, selection bias, and measurement error to identify what is needed for the model to be valid.
3. Specification of the target quantity, model specification, and model estimation	Specify whether predictors or predicted individual risk is of interest. Select predictors to be included in the model without overfitting. If overfitting is a concern, assess whether shrinkage methods should be considered.	Use the DAG to select appropriate predictors. Determine which variables should be fixed/forced when going through variable selection processes. The DAG can also inform potential sources of heterogeneity If the aim is prognostic or counterfactual prediction use appropriate variable selection methods for identification if necessary. If the aim is diagnostic or predictor strength, then use the Markov Blanket to optimize predictor selection.
4. Model performance & validation (Bias analysis)	Assess appropriateness of selected predictors through the use of calibration, discrimination, and clinical usefulness methods. Assess internal and external validity of the model	Use the DAG to evaluate external validity. Assess whether issues arise where the validity of the model may not be applicable in other populations (transportability). Use the DAG to guide quantitative bias analysis to test if selection bias, measurement error, or unmeasured confounding are an issue.
5. Interpretation of results	Interpret the results of the model taking into account: the target population, transformations, uncertainties, exclusion/inclusion criteria, and validity	Use the DAG to guide the interpretation based on whether the target population can be inferred on.
*Not explicitly included in the original modeling steps outlined by Steyerberg and Vergouwe[15]		

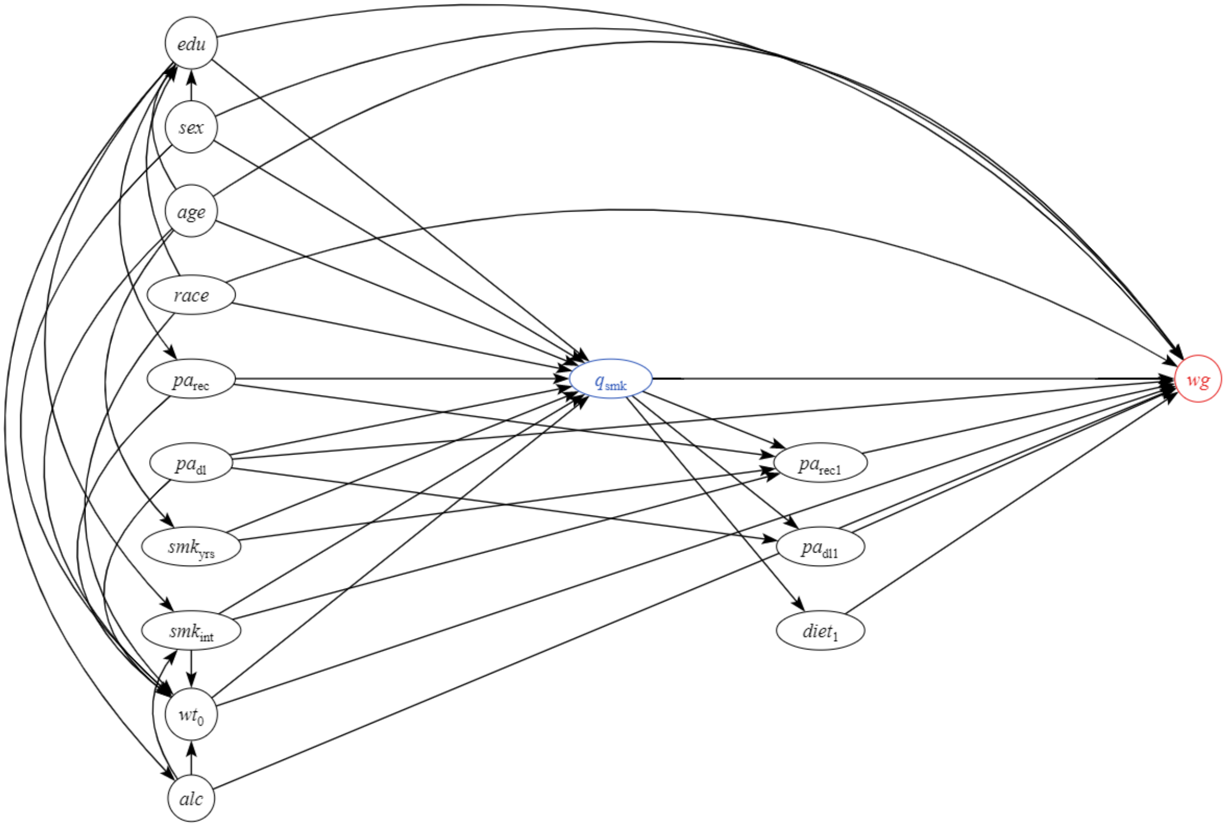

Figure 3.1: This DAG shows the relationships between smoking cessation (q_{smk}), potential confounders, potential mediators, and weight gain (*wg*). Baseline confounders include: education (*edu*), sex (*sex*), age (*age*), race (*race*), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), smoking intensity (smk_{int}), years of smoking (smk_{yrs}), weight at baseline (wt_0), and alcohol frequency (*alc*). The potential mediators are post-smoking cessation recreational physical activity (pa_{rec1}), post-smoking cessation physical activity in daily life (pa_{dl1}), and post-smoking cessation diet ($diet_1$).

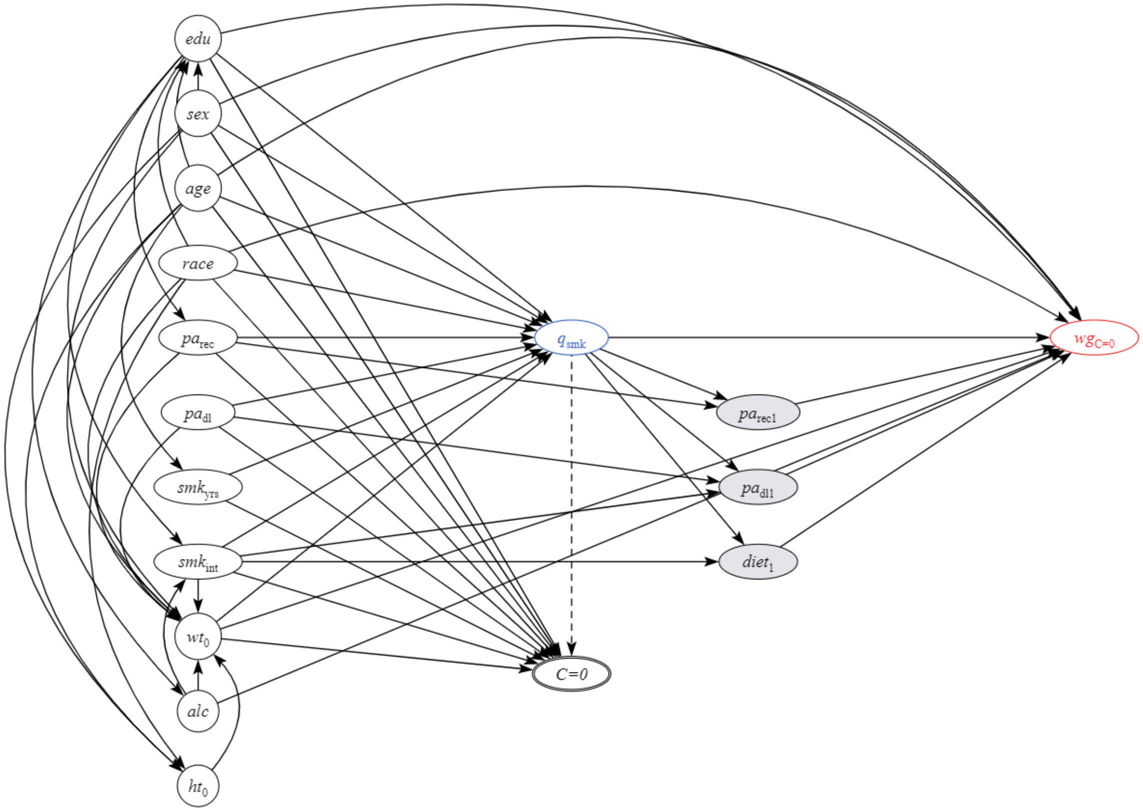

Figure 3.2: This DAG shows the relationships between potential predictors, and weight gain (*wg*) in relationship to the data. Predictors include: education (*edu*), sex (*sex*), age (*age*), race (*race*), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), smoking intensity (smk_{int}), years of smoking (smk_{yrs}), weight at baseline (wt_0), height at baseline (ht_0), alcohol frequency (*alc*), smoking cessation (q_{smk}), post-smoking cessation recreational physical activity (pa_{rec1}), post-smoking cessation physical activity in daily life (pa_{dl1}), and post-smoking cessation diet ($diet_1$). A censoring node (*C=0*) is included to account for censoring on missing data in the outcome.

Table 3.3: Frequency and means of characteristics of the study population, NHEFS 1982-1984.

Characteristic	***Total sample population (N=1,566)***
	n (%)
Quit smoking	
1: Yes	403 (25.7%)
0: No	1,163 (74.3%)
Sex	
0: Male	762 (48.7%)
1: Female	804 (51.3%)
Race	
0: White	1,360 (86.8%)
1: Black or other	206 (13.2%)
Education at baseline	
1: 8th Grade or less	291 (18.6%)
2: High School Dropout	340 (21.7%)
3: High School	637 (40.7%)
4: College Dropout	121 (7.7%)
5: College or more	177 (11.30%)
Exercise at baseline	
0: Much exercise	300 (19.2%)
1: Moderate exercise	661 (42.2%)
2: Little or no exercise	605 (38.6%)
Daily activity at baseline	
0: Very active	702 (44.8%)
1: Moderately active	715 (45.7%)
2: Inactive	149 (9.5%)
Alcohol frequency at baseline	
0: Almost every day	325 (20.8%)
1: 2-3 times/week	219 (14.0%)
2: 1-4 times/month	494 (31.5%)
3: < 12 times/year	328 (20.9%)
4: No alcohol last year	195 (12.5%)
5: Unknown	5 (0.3%)
	Mean (95% CI)
Age (years)	43.7 (43.5, 43.9)
Change in weight (kg)	2.6 (2.5, 2.8)
Weight at baseline (kg)	70.8 (70.6, 71.1)
Smoking intensity at baseline (cigarettes/day)	20.5 (20.3, 20.7)
Years of smoking	24.6 (24.4, 24.8)

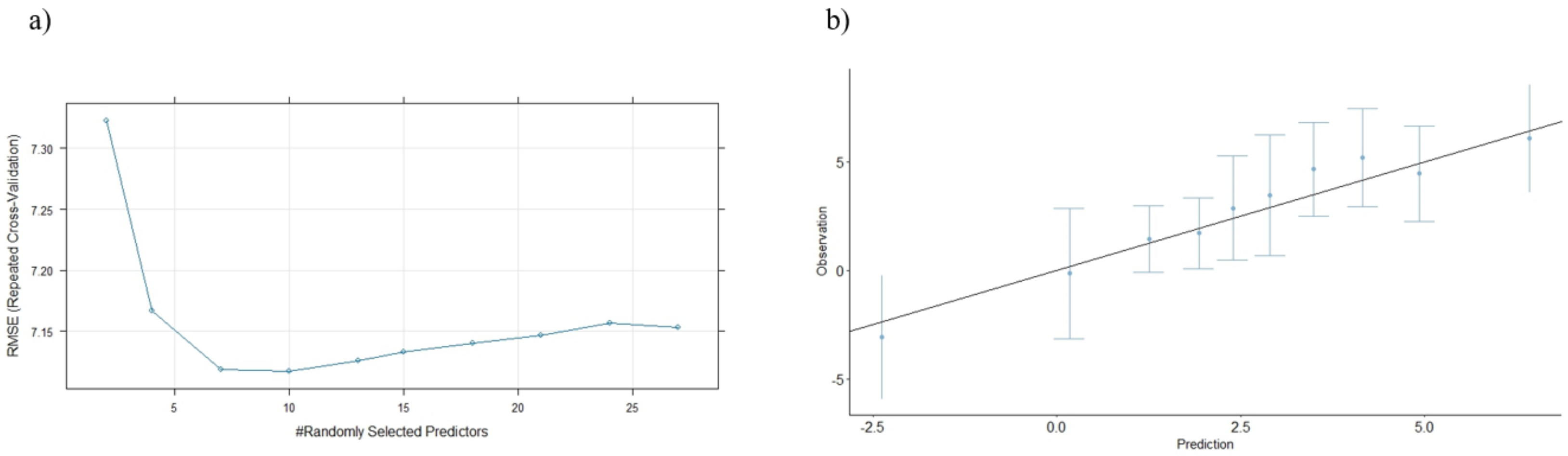

Figure 3.3: Plots for model specification, performance, and validation. **a)** Plot of the root mean squared error (RMSE) after repeated 5-fold cross-validation with 5 repetitions and tuned to 10 variations of mtry (i.e. the number of randomly selected predictors) for the training data **b)** The calibration plot comparing predicted weight gain to observed weight gain in the testing data.

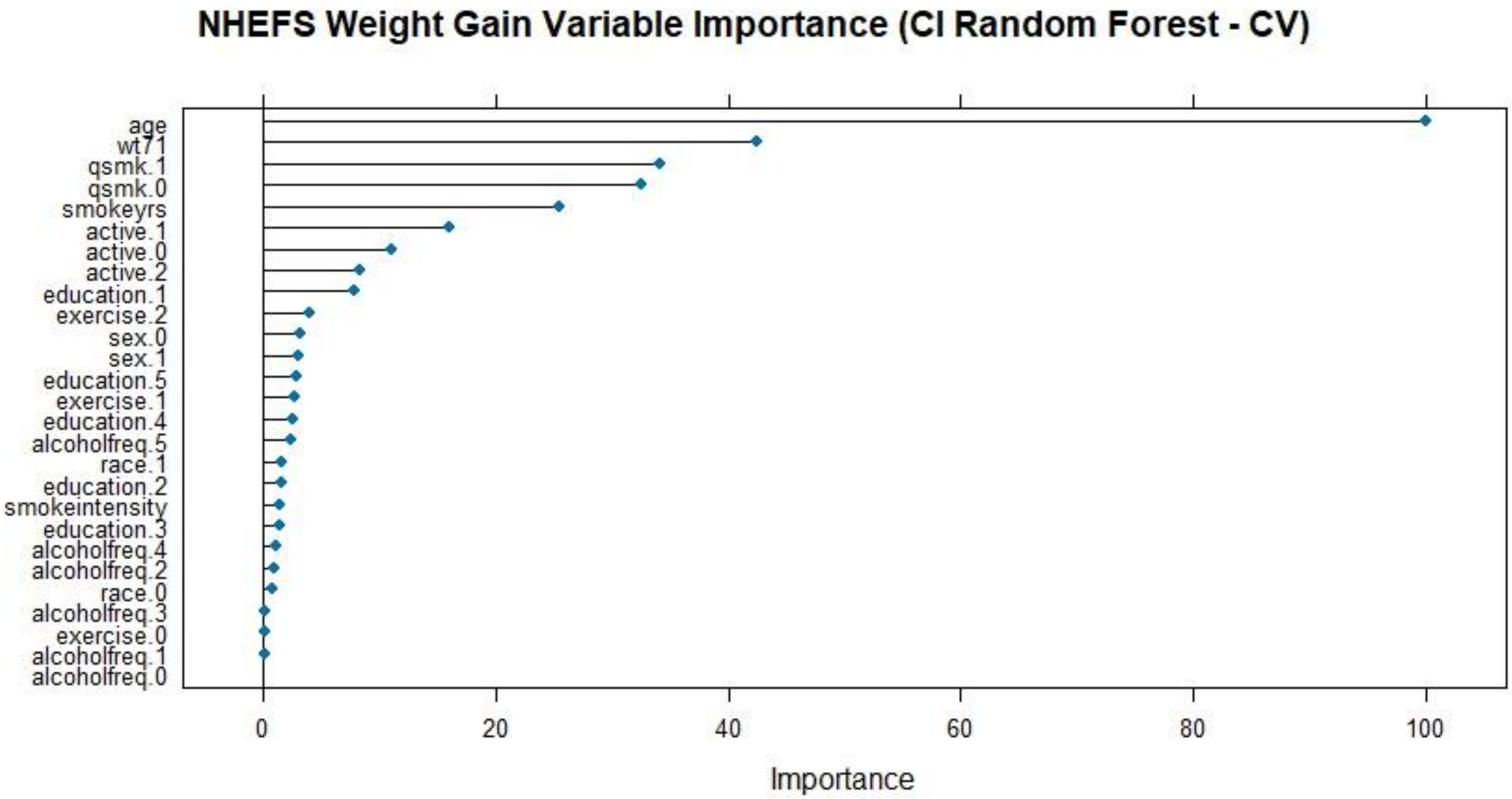

Figure 3.4: Variable importance plot from a conditional inference random forest to predict weight gain among smokers, NHEFS 1982-1984.

Chapter 4 DAGging out causal aims

4.1 Abstract

Causal inference is a cornerstone of epidemiologic research. Identifying and quantifying the mechanisms by which disease occurs is a primary goal in public health for prevention. Students are often taught epidemiology from a causal perspective and directed acyclic graphs (DAGs) are used to clarify assumptions and illustrate causal questions to inform study design and statistical analysis. Although DAGs have become an increasingly popular method for causal inference, the use of DAGs in applied research remains low. Here, we aim to provide guidance on the use of directed acyclic graphs (DAGs) while navigating the causal roadmap. We apply this adapted framework to assess the average causal effect of smoking cessation on weight gain in the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS). We discuss how the DAG can evolve and be augmented in response to the data and analytic decisions. We also demonstrate the importance of using the DAG to identify and address sources of confounding, selection bias, information bias, and missing data to prevent bias and communicate uncertainties.

4.2 Introduction

4.2.1 Directed Acyclic Graphs (DAGs) and Causal Inference

Directed Acyclic Graphs (DAGs) are used in epidemiology and clinical research to clarify assumptions and illustrate causal questions to inform study design and statistical analysis.[1–4] DAGs are generally considered to be *causal* diagrams. Causal diagrams are visual tools used to depict these types of relationships. DAGs provide a method to visualize and check dependencies among variables for model specification.[2] As such, they are instrumental in assessing whether certain conditions for identifiability have been met. Additionally, DAGs allow researchers to investigate different scenarios through the manipulation of variables. Thus, simulating counterfactuals and interventions becomes possible to explore.[2] DAGs are adaptable and flexible due to their non-parametric nature. Their mathematical foundations have been previously described.[1,5,6] Several resources have been published that provide the basics of DAGs.[1–5,9,10] Furthermore, Ferguson et al have provided guidelines on the synthesis of evidence and the construction of DAGs to provide a systematic method of the development of DAGs for causal inference.[12]

Although DAGs have become an increasingly popular method for causal inference, the use of DAGs in applied research remains low.[7] A recent study found that around 40% of respondents did not use DAGs, and the most common reason given was that they did not know how to use them.[7] Accessibility to relevant training resources may be a barrier to more widespread use.[7] Specifically, the use and reporting of DAGs vary in applied health research.[8] Even when DAGs are reported in these studies, most fail to report how adjustment sets were derived for estimates that are provided, including those for the primary analysis of interest.[8] Confusion or disagreement on the rules and assumptions of DAGs may have contributed to their limited uptake as a systematic method to construct models.[7,8] Additionally, when researchers use DAGs, it is in the limited context of presenting knowledge and assumptions and selecting deconfounders. However, DAGs can evolve and be augmented to present the data-generating process. The causal roadmap developed by Petersen and van der Laan provides researchers with a framework to address causal questions.[16] This study aims to describe how DAGs can be used at each stage of the causal roadmap as well as to demonstrate the use of DAGs and the causal roadmap through the application to data taken from *Causal Inference: What If*.[4]

4.3 Navigating the causal roadmap with DAGs

4.3.1 Overview of the causal roadmap

The causal roadmap that Petersen and van der Laan developed includes seven steps to address causal questions.[16] These steps were further explored with application by Balzer et al. to provide guidance on the use of the causal roadmap.[79] Here, we explore the use of DAGs while navigating the causal roadmap with an application that was originally used in *Causal Inference: What if*.[4] The causal roadmap steps and the potential use of DAGs at each step have been outlined in **Table 4.1**. Some steps were combined for simplicity however, the considerations remain the same as if they were separated. We also included a step that is specific to conducting sensitivity/bias analysis to ensure that the sensitivity of the estimate to potential biases or assumption violations is fully explored. In the following sections, we will describe in detail how DAGs can be augmented and evolve to be useful at each step of this causal roadmap.

4.3.2 Step 0: Specify the research question

Before embarking on this journey, we need to define the research question of interest. Once the research question has been defined, we can determine whether the question is causal in nature. Part of the process of defining the research question includes specifying what the target population is. It is important to understand what population we would like to make inferences about. Later, this will help identify the measure of effect of interest, whether we are interested in making inferences on the total population, the treated population, or the untreated population. Additionally, we will need to define the outcome and exposure/treatment of interest. This step is a preparation step that helps us to begin to build our DAG.

4.3.2.1 Step 0: Application

For the purposes of demonstration, we will explore the causal relationship between smoking cessation and weight gain as was done in *Causal Inference: What if*.[4] Smoking is known to cause harm to nearly every organ in the body and as such, those who quit smoking experience better health outcomes than those who do not.[41,42] However, a potential barrier to smoking cessation is concern over subsequent weight gain.[43–45] Thus, we may wish to know the average total effect of smoking cessation on weight gain to potentially alleviate these concerns. As such, the aim of our study is to quantify the total average causal effect of smoking cessation on weight gain. Our target population is the analytic sample which consists of smokers with complete data on covariates from the civilian adult population of Americans ages 25-74 years selected in 1971-1975 and followed-up in 1982-1984.

4.3.3 Step 1: Specify knowledge about the system

When addressing a causal aim, there is generally some information available on the relationships between variables. A literature review should be conducted to establish what prior knowledge exists about the outcome and exposure of interest as well as any potential covariates. This information can then be synthesized and displayed using a DAG to visualize the relationships between variables. By specifying prior knowledge, we can evaluate what relationships need to be accounted for to remove bias and ensure identifiability. If data has not been collected yet, specifying the DAG may aid in data collection and inform study design. A protocol developed by Ferguson et al to standardize the development of DAGs and encode this background knowledge provides a step-by-step process to map, translate, and integrate the information into a DAG.[12]

4.3.3.1 Step 1: Application

Figure 4.1 depicts a DAG that illustrates the potential causal relationships between smoking cessation status, potential confounders, potential mediators, and weight gain. In our example, our outcome of interest is weight gain at follow-up (wg). Our exposure/treatment of interest is smoking cessation between baseline and follow-up (q_{smk}). Technically, this treatment is time-varying since smoking cessation can take place at any point between baseline and follow-up as it is strictly voluntary and should be treated as such.[4] However, in our example, we will ignore the time-varying aspect of the treatment in favor of simplicity as was done in Chapter 12 of *Causal Inference: What if.*[4] Baseline covariates such as education (edu), sex (sex), age (age), race ($race$), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), years of smoking (smk_{yrs}), smoking intensity (smk_{int}), weight in kilograms (kg) (wt_0), and alcohol frequency (alc) were evaluated as potential confounders of the relationship between smoking cessation status and weight gain at follow-up. We also acknowledge that indirect causal paths between smoking cessation and weight gain exist where it can be mediated by post-treatment recreational and daily life physical activity (*pa_{rec1}* and *pa_{dl1}*, respectively) as well as post-treatment diet (*$diet_1$*).

4.3.4 Step 2: Link the observed data to the DAG

In the second step of the causal roadmap, we would need to link the observable data to the DAG. By doing so, any potential threats to internal validity can then be assessed.[80] These potential threats include (but are not limited to): 1) ambiguous temporality; 2) sources of measurement error; 3) selection; and 4) missing data. If the temporality of variables cannot be established, reverse causation may be of concern as it cannot be ruled out. The directionality of the arrows in DAGs is intended to establish a temporal precedent to identify this potential issue.[80] Similarly variables that occur in concurrence with the exposure and can also be viewed as an ancestor of the outcome may also be a threat to validity. As such, we should identify which variables (if any) are unmeasured and potentially introduce proxy variables that are being used to indirectly assess these unmeasured variables in their place if necessary. Including potential sources of measurement error in the DAG may also be necessary to assess whether misclassification bias may be an issue. Joint effects may also contribute to measurement error and should be considered when assessing potential bias and the use of proxies.[80] Such joint effects can induce collider bias and should be carefully considered.[80] Additionally, we should also identify the relationship between the target and study populations. To do so, we will need to describe the criteria for inclusion or exclusion in the study population. These criteria can be displayed in the DAG through the inclusion of a selection variable. Similarly, it may be important to include if and how study participants were lost to follow-up. Visualizing the study population in a DAG can help further illuminate potential sources of selection bias and/or transportability.[4,34,53,58] Lastly, if missing data is an issue, nodes can be included that describe patterns in missing data to assess their level of randomness so that an appropriate method for adjusting for the missing data can be identified.[29,72,73,81]

4.3.4.1 Step 2: Application

In our example, we use the National Health and Nutrition Examination Survey I (NHANES-I) Epidemiologic Follow-up Study (NHEFS) as used by Hernán and Robins.[4,17,18] NHANES-I was conducted in 1971-1975 and the NHEFS was conducted in 1982-1984 as a follow-up. Participants in NHEFS included persons 25-74 years of age who completed the medical examination at baseline for NHANES-I (n = 14,407).[17] The sampling procedure for NHANES-I is described in detail elsewhere.[46,47,82] Certain groups of individuals were oversampled to ensure representativeness in the survey and thus, sampling weights as well as cluster and strata information are provided with the data to account for this oversampling. If our target population was the general non-institutionalized smoking population of the contiguous United States, we may need to use these weights in order to make inference on the target population. Similarly, as shown in the DAG in **Figure 4.2**, our analytic sample is a result of restricting on individuals with known weight, sex, age, race, height, smoking intensity, and alcohol use at baseline and censoring those with missing weight measurement at follow-up (n = 1,566).[4] In this DAG, a censoring node ($C=0$) has been added to reflect this censoring. Additionally, although smoking cessation status was not directly censored on, it may still be associated with the probability of being censored. As pointed out by Hernán and Robins, the censoring of those with missing weight measurement at follow-up may have induced selection bias by conditioning on a post-treatment event (i.e., participation in the 1982 follow-up). Thus, if smoking cessation is associated with the probability for censoring, then selection bias may be induced because smoking cessation is actually a time-varying treatment.[4] This censoring would result in selection bias if our target population went beyond the analytic sample itself. Thus, we would need to account for this selection bias as well in order to make inferences beyond the analytic sample.[48,53,54,59] However, in our example, our target population is the analytic sample itself. To reflect this in **Figure 4.2**, weight gain is in terms of the uncensored ($wg_{C=0}$).

The outcome, exposure, and covariate definitions for our example are consistent with those from Chapter 12 of *Causal Inference: What If*.[4] Our outcome of interest is weight gain at follow-up which is a continuous variable that reflects the weight in kg at follow-up minus the weight in kg at baseline. An individual was considered to have quit smoking between baseline and follow-up if they reported having quit smoking at follow-up, else they were considered to have not quit smoking (1: Yes; 0: No). Age (years), years of smoking, smoking intensity (cigarettes/day), and height in centimeters were also continuous variables and measured at baseline. Sex (male vs female), race (White vs Black or other), education (1: 8th grade or less; 2: high school dropout; 3: high school; 4: college dropout; 5: college or more), recreational physical activity (0: much exercise; 1: moderate exercise; 2: little or no exercise), daily life physical activity (0: very active; 1: moderately active; 2: inactive), and alcohol frequency (0: almost every day; 1: 2-3 times/week; 2: 1-4 times/month; 3: < 12 times/year; 4: no alcohol in the last year; 5: unknown) were all categorical and also measured at baseline.

4.3.5 Step 3: Specify the target causal quantity and assess identifiability

As described in **Box 4.1**, we must ensure that all variables are well-defined to properly assess identifiability. This is especially true in the case of the intervention of interest. It is important that we fully define the intervention of interest and its counterfactual quantity to properly measure the causal effect. Intervening on a variable changes the system it is operating in by fixing its value.[5] Therefore, it can be useful to visualize how the system is affected after the intervention has taken place in a post-intervention DAG. Specifically, the post-intervention DAG is derived from the pre-intervention DAG and are intended to show the consequences of our actions. For instance, if we intervene on treatment then the arrows going into the treatment are removed. As a result, the post-intervention DAG helps illuminate whether the intervention is complete on its own or whether other conditions are necessary to measure the causal effect and meet the conditional exchangeability condition. Other considerations for the causal effect of interest include whether the question calls for the direct, indirect, or total effect.[83,84] Reflecting on whether selection bias or transportability was an issue in step 2, we can assess whether the chosen measure of effect can be estimated and inferences of the target population can be made or identify if another measure of effect may be more appropriate. Similarly, it is important to know if the effect of interest is a joint effect or requires accounting for effect modification.[65] Several studies have proposed how to use DAGs to depict interaction and effect modification through augmentation. [66,67,85–89] Understanding which effect is of interest will dictate what is required for identifiability.[83,84,90–93]

We also need to evaluate whether the observed data and the intervention are sufficient to answer the question of interest. Again, we refer to **Box 4.1** to assess whether the conditions for identifiability have been met or whether further adjustments are needed. Unlike intervening on a variable, however, adjustment does not alter the system. Instead, when a researcher adjusts or conditions on a variable, we are only narrowing the focus to a certain subset of cases.[5] Looking at the post-intervention graph, we can use an appropriate identification framework to select variables to be included in the model for adjustment. Such identification frameworks include Pearl's backdoor criterion or the front-door criterion should unmeasured confounding be an issue and an intermediate variable is available.[5] Methods for selection of deconfounders have been previously proposed elsewhere.[5,94,95] It may be the case that several minimally sufficient sets are available to control for confounding by using different criteria. In which case, to choose the best set, we would need to consider which set is less prone to measurement error or missing data. If competing causal structures are an issue, minimally sufficient sets from all causal structures should be considered. Furthermore, additional considerations may need to be addressed to assess the effect of interest.

4.3.5.1 Step 3: Application

Our target causal effect is the average causal effect of smoking cessation on weight gain. Specifically, we are interested in quantifying the average total effect of smoking cessation between baseline and follow-up on weight gain at follow-up. We will use the backdoor criterion to identify confounders and assess identifiability. Resources such as DAGitty and Causal Fusion are useful tools to assess identifiability based on the queried DAG.[96–98] To assess whether the conditional exchangeability assumption holds, we check the DAG in **Figure 4.2** (i.e., the pre-intervention DAG) to assess whether there are any confounders. The DAG indicates that there is a single minimally sufficient set for control of confounding. This set includes the baseline confounders education, sex, age, race, recreational physical activity, physical activity in daily life, years of smoking, smoking intensity and weight. This set of confounders is consistent with those used in Chapter 12 of *Causal Inference: What if*.[4] This is clarified in the post-intervention DAG (**Figure 4.3**) when we intervene on smoking cessation status. We can further assess this by evaluating the distribution of these potential confounders by smoking cessation status (**Table 4.2**) where we do notice a difference in distribution of baseline confounders by smoking cessation status. However, we can also see that there is still potential selection bias that was induced by censoring on missing data. If smoking cessation is associated with censoring, selection bias could be an issue. This would preclude identifiability of the average causal effect of smoking cessation on weight gain. Sensitivity/bias analysis will be necessary to quantify the magnitude of this bias. Another source of potential bias is measurement error which in this example we assume does not exist for simplicity however, potential sources of measurement error or misclassification bias should be evaluated further in any causal study to ensure that identifiability holds. Hernán and Robins also discuss the empirical violation of the positivity assumption since some strata exist that have a probability of smoking cessation equal to zero.[4] This is not surprising since the risk of a violation of the positivity assumption is higher with higher dimensional data.[79] In this example, interference is not an issue since the weight gain of one individual does not affect the weight gain of another.

4.3.6 Step 4: Commit to a statistical model and estimand

Once we have assessed what would be required for identifiability of the causal quantity, it is time to commit to a statistical model and estimand. At this point, we should have enough information to decide whether the effect of interest is identifiable. It may be the case that identifiability of the estimand is not possible. One option is to declare the unidentifiability of the estimate and move on. Another option is to select a different estimand that is similar but requires additional assumptions. Petersen and van der Laan differentiate between convenience-based assumptions and knowledge-based assumptions to select an appropriate statistical model and estimand.[16,79] Unfortunately, it is not uncommon for measured variables to be insufficient to control for confounding and instead should be assessed through quantitative bias analysis.[53,54,57,60,99] Assumptions that arose from a literature review or expert input are generally categorized as knowledge-based assumptions and tend to lie closer to the truth; however, often to improve identifiability, additional assumptions are made that may not be based on real knowledge and are thus categorized as convenience-based assumptions. It is important to differentiate between these assumptions to select an appropriate model and estimand. When selecting variables for models for causal inference, the primary concern is to ensure conditional exchangeability.[4] Thus, a DAG depicting only knowledge-based assumptions and another DAG that includes convenience-based assumptions should be compared. Statistical models should be selected based solely on knowledge-based assumptions to ensure that it contains the truth.[16,79] However, it may be the case that using convenience-based assumptions are unavoidable. Thus, we should minimize the number of convenience-based assumptions used to select an estimand, if possible.

4.3.6.1 Step 4: Application

In the previous step, we noticed that there are a few assumptions violated that may preclude identifiability. Namely, the violation of the positivity assumption. According to Hernán and Robins, the violation of the positivity assumption in this case was a random violation.[4] Random violations of the positivity assumption occur when the sample is finite such that when we stratify on several confounders, we are bound to find zero cells even when the probability of treatment is not actually zero.[4] Additionally, we identified several potential confounders that need to be adjusted for in order to ensure conditional exchangeability. To remove the influence of these confounders on smoking cessation status, we will use stabilized inverse probability of treatment weights (IPTW). By doing so, we will be able to smooth over the random violation of the positivity assumption by modeling the probability of treatment in strata with random zeroes using the data from the individuals in other strata.[4] This is further clarified in **Figure 4.3** which indicates that by applying the IPTWs, we remove the arrows going into smoking-cessation status. The weight we construct will have the following form:

$$IPTW_{qsmk=1} = \frac{\Pr(qsmk = 1)}{\Pr(qsmk = 1|Z_0)}$$

and

$$IPTW_{qsmk=0} = \frac{1 - \Pr(qsmk = 1)}{1 - \Pr(qsmk = 1|Z_0)}$$

where we let Z_0 represent the minimally sufficient set of baseline confounders age, sex, race, education, recreational physical activity, daily life physical activity, smoking intensity, years of smoking, and weight. We assume a parabolic relationship between continuous confounders and the smoking cessation status (a dichotomous variable). Thus, quadratic terms for continuous variables (i.e. age, weight at baseline, smoking intensity, and years of smoking) were included in the model of treatment. We use bootstrapping methods to calculate 95% confidence intervals. Fortunately, since we are assessing the total effect of smoking cessation on weight gain, we do not require the measurement of the mediators to ensure identifiability. However, if we were interested in the direct effect of smoking cessation on weight gain, we would require additional data. Our knowledge-based assumptions are sufficient for conditional exchangeability to hold. Selection bias may still be an issue depending on our target population. Again, we may wish to explore the degree of selection bias with a sensitivity/bias analysis. All analyses were conducted using SAS® version 9.4 statistical software and the seed for bootstrapping methods was 1234567.

4.3.7 Step 5: Estimation

We now can estimate our causal effect of interest. We take what we learned in steps 0-4 to estimate our causal effect. Additionally, any analytic decisions made during this step should be documented in the DAG to guide interpretation later on. If interaction terms were needed or transformations were required, we may wish to include these in the DAG to assess how using them will affect our inference. This is especially true if the transformation is used as a proxy for the variable.

4.3.7.1 Step 5: Application

In our example, we assumed a parabolic relationship between continuous baseline covariates and smoking cessation status; however, the squared terms were not used as a proxy for the continuous variable. Thus, our pre- and post-intervention DAGs still hold (**Figures 4.2** & **4.3**).

4.3.8 Step 6: Sensitivity/bias analysis

It is an unfortunate fact that any type of research contains some level of uncertainty. Often when conducting causal studies, we would like to assume that all error is random. However, to do so would be unrealistic as systematic error does exist and can heavily bias our estimate. In fact, an untested biased estimate in public health can have severe consequences when it is mistaken for the truth as it has the potential to influence policy and/or clinical decisions. Thus, bias analysis is a necessary step to ensure that sources of potential bias have been acknowledged and/or addressed.[9,53,57,60,61,100–102] Transparency of the uncertainties involved in the estimation of the measure of effect is essential to advance knowledge for public health research and practice.[100] To guide our sensitivity analysis, we use the DAGs we created in the previous step. If competing causal diagrams have been proposed, we would need to assess the magnitude of bias in each causal system. In some cases, proper sensitivity analyses would require external data to quantify the magnitude of the bias since the data could be insufficient to account for unmeasured covariates or the assumptions themselves are untestable.[60,79,99,102] Any potential lingering biases or uncertainties that limit our ability to address our research question should be quantitatively explored. Whether we are concerned about unmeasured confounders, selection bias, measurement error, or violations of the conditions of identifiability, sensitivity analysis will enhance transparency of the limitations of the estimate so that future research can aim to avoid these pitfalls.

4.3.8.1 Step 6: Application

To assess the potential selection bias induced by censoring on missing weight measurement at follow-up, we will test the sensitivity of the estimate attained by using the IPTW weights alone in comparison to an estimate of the mean difference attained by a model where we use these same IPTW weight as well as inverse probability of censoring weights (IPCW). The stabilized IPCW weights we use will have the following form:

$$IPCW = \frac{\Pr(C = 0|qsmk)}{\Pr(C = 0|\ qsmk, Z_0)}$$

where the full weight will combine the IPTW and IPCW to emulate random censoring with respect to the baseline confounders in Z_0. To combine the weights, we multiply the IPTW by the IPCW ($IPTCW = IPTW * IPCW$). **Figure 4.4** is a simplified post-intervention DAG showing that by using these IPTCWs we are able to remove the arrow from Z_0 to censoring. Essentially, by doing so, we are assessing the mean weight gain if everyone had quit smoking and nobody had been censored, a joint effect.[4]

4.3.9 Step 7: Interpretation

Here, we take what we learned from steps 0-6 to guide the interpretation of our results. At this point, we should have enough information to communicate any lingering uncertainties that remain. We also assess our ability to make inferences on the target population. The statistical interpretation of our estimand may not change, however, the causal interpretation can vary depending on the limitations of the data and the degree of bias. Thus, the results and implications from any sensitivity/bias analysis conducted should be clearly communicated to prevent misuse of the estimate in future research.

4.3.9.1 Step 7: Application

The total number of participants included in the analytic sample was 1,629 after initial restriction to complete baseline covariate measurements and 1,566 after another 63 were censored due to missing weight measurement at follow-up. Due to the potential selection bias and oversampling during survey design, we refrain from extending the inference of our results beyond the analytic sample. Initial descriptive analysis of the analytic sample to assess conditional exchangeability showed that those who quit smoking varied slightly from those who did not quit smoking especially in terms of baseline covariates sex, age, education, weight, smoking intensity, and years of smoking (**Table 4.2**). The estimated stabilized IPTWs ranged from 0.33 to 4.21 and had a mean of 1.00. Additionally, the estimated stabilized IPCWs ranged from 0.94 to 1.72 and had a mean of 1.00. When we estimated the combined IPTW*IPCW, the weights ranged from 0.35 to 4.09 with a mean of 1.00. Using the stabilized IPTWs to adjust the model for baseline confounders age, sex, race, education, smoking intensity, years of smoking, recreational physical activity, physical activity in daily life, and baseline weight, we estimated that on average, those who quit smoking gained 3.52 kg more than those who did not quit smoking at follow-up (Bootstrap 95% CI: [2.45,4.49]) (**Table 4.3**). Similarly, when we used the combined IPTW*IPCW, we estimated that on average, quitting smoking increased weight by 3.50 kg (Bootstrap 95% CI: [2.43, 4.44]). Our estimates are consistent with those estimated by Hernán and Robins.[4] The similarity between these two estimates can indicate that selection bias is either minimal or could not be removed.

4.4 Conclusion

Causal aims are of chief concern in epidemiology to investigate and explain the mechanisms by which disease occurs. DAGs are traditionally used in causal studies to inform analysis and data collection, communicate assumptions, assess identifiability, detect the presence of bias, and guide interpretation. In our paper, we reiterate the value of DAGs to address causal aims while navigating the causal roadmap by demonstrating their use at each step. The use of DAGs throughout the process can help clarify assumptions and guide interpretation to communicate uncertainties that could not be resolved. Though our example is simple, we demonstrate some of the considerations that should be made when navigating the causal roadmap. However, DAGs heavily rely on the assumption of faithfulness and are therefore limited in their usefulness if faithfulness cannot be assumed. Nevertheless, DAGs are flexible tools that can be augmented and evolve. When addressing a causal aim, DAGs will evolve to better assess changes to the causal structure as a result of data collection procedures, lost-to-follow up, missing data, data fusion, and analytic decisions.

4.5 Appendix

Table 4.1: Outline of steps of the causal roadmap and the use of Directed Acyclic Graphs (DAGs)

Step	***Description***	***Use of DAGs***
0. Define the research question	Define the research question. Define the target population.	Not necessarily applicable - this step is preparation for DAG creation and scope of study. A basic DAG consisting of the outcome and a primary covariate (if applicable) of interest can be drawn but is not necessary.
1. Specify knowledge about the system	Describe background knowledge with respect to the question under study	Display background knowledge. Encode relationships between variables in the DAG.
2. Link the observed data to the DAG	Identify which variables have been measured. Describe selection/inclusion criteria. Link the observable data to the causal model. Define the sample/study population.	Identify sources of confounding, selection bias, and/or measurement error. Assess issues with missing data. Augment the DAG to specify what data is observable vs unobservable
3. Specification of the target causal quantity & assess identifiability	Define the intervention/exposure variables and the type of intervention. Identify the measure of effect of interest. Specify the model. Assess identifiability conditions and whether additional assumptions are needed for identification. Specify an identification framework.	Use pre- and post-intervention DAG to define the intervention and its potential effects. Use the specified identification framework to select variables for the model. Assess whether conditions for identifiability have been met.
4. Commit to a statistical model and estimand	Identify the causal effect of interest and the corresponding estimand. Assess whether knowledge is sufficient to identify the causal effect using the specified statistical model.	Augment the DAG to include post-intervention and backdoor path identification methods. Address additional concerns that may arise due to the statistical model of choice.

Step	***Description***	***Use of DAGs***
5. Estimation	Estimate the causal effect.	Augment the DAG to include any analytic decisions made to guide the interpretation later
6. Sensitivity/Bias analysis*	Test sensitivity of the causal effect due to assumptions made or potential bias	Guide the sensitivity or bias analysis using a DAG to identify sources of uncontrolled confounding, selection bias, or measurement error/information bias
7. Interpretation	Interpret the results appropriately accounting for any biases or assumption violations that may still exist preventing effect identification or inference on the target population	Guide the interpretation of the selected estimate based on what was done during the analysis and whether the target population can be assessed.
*Not included in the original roadmap proposed by Petersen and van der Laan[16]		

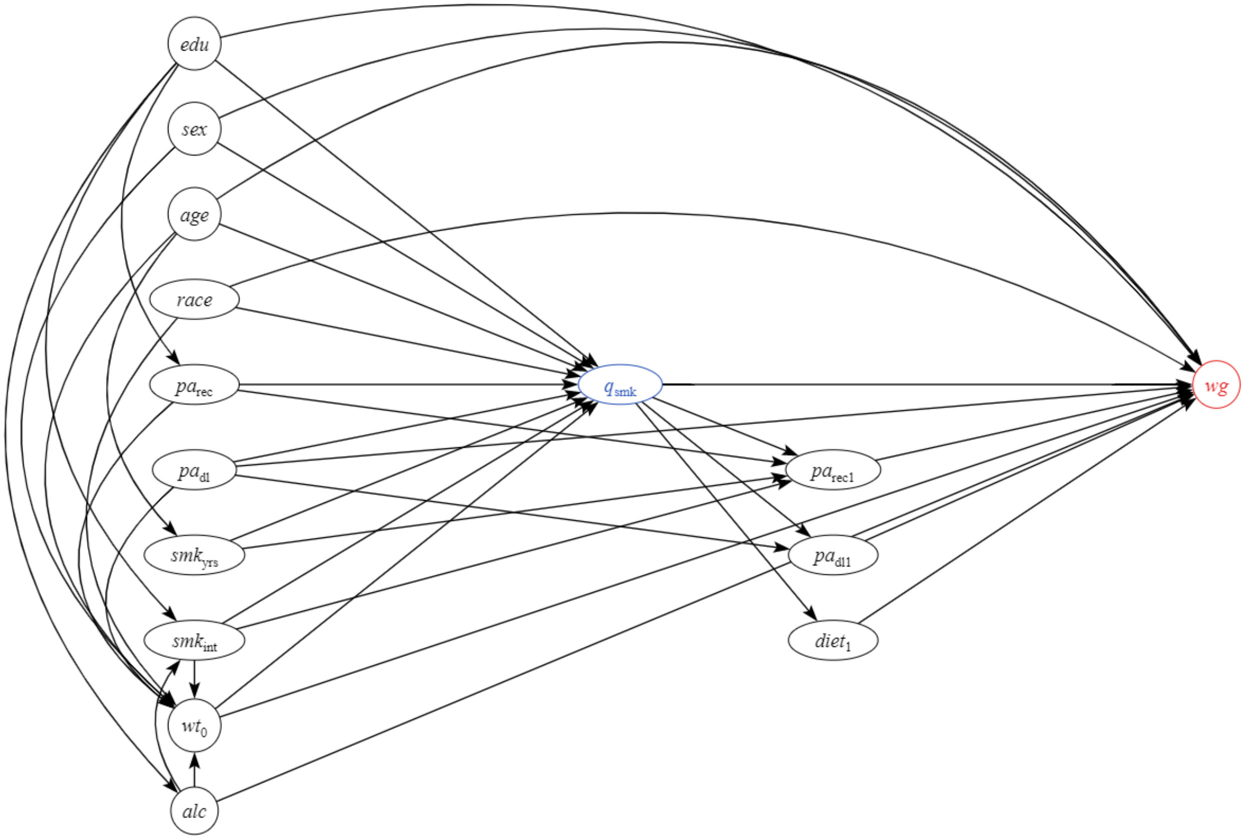

Figure 4.1: This DAG shows the relationships between smoking cessation (q_{smk}), potential confounders, potential mediators, and weight gain (*wg*). Baseline confounders include: education (*edu*), sex (*sex*), age (*age*), race (*race*), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), smoking intensity (smk_{int}), years of smoking (smk_{yrs}), weight at baseline (wt_0), and alcohol frequency (*alc*). The potential mediators are post-smoking cessation recreational physical activity (pa_{rec1}), post-smoking cessation physical activity in daily life (pa_{dl1}), and post-smoking cessation diet ($diet_1$).

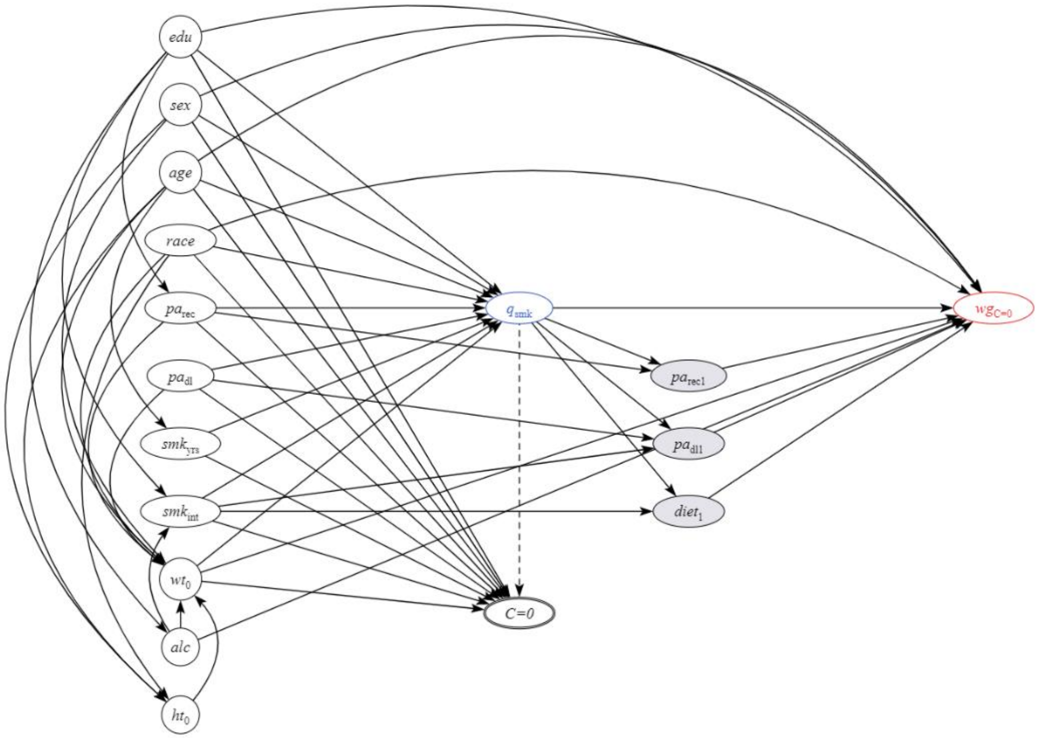

Figure 4.2: This DAG shows the relationships between smoking cessation (q_{smk}), potential confounders, potential mediators, and weight gain (*wg*) in the context of the data. Baseline confounders include: education (*edu*), sex (*sex*), age (*age*), race (*race*), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), smoking intensity (smk_{int}), years of smoking (smk_{yrs}), weight at baseline (wt_0), and alcohol frequency (*alc*). The potential unmeasured mediators are post-smoking cessation recreational physical activity (pa_{rec1}), post-smoking cessation physical activity in daily life (pa_{dl1}), and post-smoking cessation diet ($diet_1$). A censoring node (*C=0*) is included to illustrate the censoring process undertaken when censoring on missing data in baseline and follow-up weight, height (ht_0), alcohol frequency, smoking intensity, age, sex, race, and education.

Box 4.1: Conditions of identifiability

1. *Well-defined variables* - Every variable is properly named and measured

2. *Conditional exchangeability* - The potential outcome of Y had X been set to x is independent of X given a set of confounders S

3. *Positivity* - For every non-zero probability of a set of confounders, S, and exposure, X, there is a greater than zero probability of X given S (i.e. all values of the exposure, X, must be possible for everyone under study)

4. *Consistency* - For those X=x, the potential outcome of Y had X=x is their observed Y

5. *Interference* - No spill-over effects or ties between units. An individual's outcome is not dependent on another individual's outcome or an appropriate model is used to account for interference

6. *No other sources of bias* - No selection bias, dependent measurement error, or misrepresentation of the data

7. *No model misspecification* - The statistical model used sufficiently accounts for existing biases and no new biases are introduced

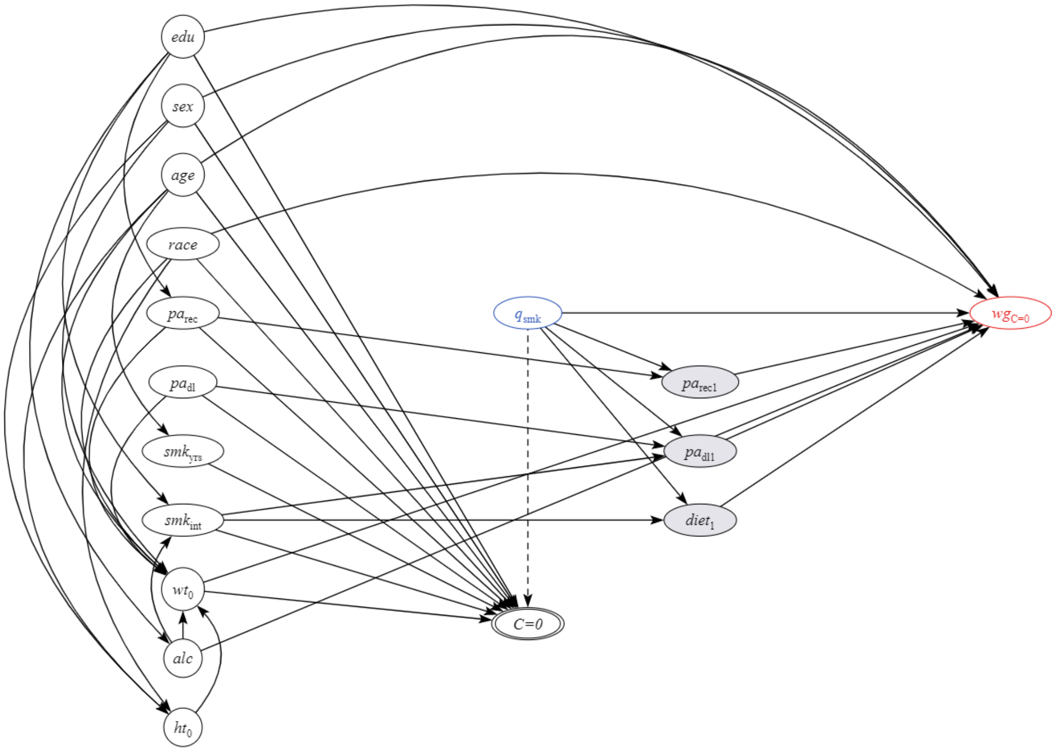

Figure 4.3: A post-intervention DAG that shows the relationships between smoking cessation (q_{smk}), potential confounders, potential mediators, and weight gain (*wg*) in the context of the data after intervening on smoking cessation. Baseline confounders include: education (*edu*), sex (*sex*), age (*age*), race (*race*), recreational physical activity (pa_{rec}), physical activity in daily life (pa_{dl}), smoking intensity (smk_{int}), years of smoking (smk_{yrs}), weight at baseline (wt_0), and alcohol frequency (*alc*). The potential unmeasured mediators are post-smoking cessation recreational physical activity (pa_{rec1}), post-smoking cessation physical activity in daily life (pa_{dl1}), and post-smoking cessation diet ($diet_1$). A censoring node (*C=0*) is included to illustrate the censoring process undertaken when censoring on missing data in the outcome. Arrows into the censoring node indicate differential lost to follow-up by baseline confounders and possible association with smoking cessation status

Table 4.2: Frequency and means of baseline characteristics for those with significant weight gain and. those without significant weight gain, NHEFS 1982-1984.

Characteristic	***Quit smoking (N=428)***	***Did not quit smoking (N=1,201)***
	n (%)	*n (%)*
Sex		
Male	237 (55.37%)	562 (46.79%)
Female	191 (44.63%)	639 (53.21%)
Race		
White	390 (91.12%)	1,024 (85.26%)
Black or other	38 (8.88%)	177 (14.74%)
Education		
8th Grade or less	93 (21.73%)	218 (18.15%)
High School Dropout	78 (18.22%)	273 (22.73%)
High School	164 (38.32%)	495 (41.22%)
College Dropout	30 (7.01%)	96 (7.99%)
College or more	63 (14.72%)	119 (9.91%)
Exercise		
Much exercise	70 (16.36%)	247 (20.57%)
Moderate exercise	181 (42.29%)	496 (41.30%)
Little or no exercise	177 (41.36%)	458 (38.13%)
Daily activity		
Very active	182 (42.52%)	547 (45.55%)
Moderately active	198 (46.26%)	540 (44.96%)
Inactive	48 (11.21%)	114 (9.49%)
Alcohol frequency		
Almost every day	89 (20.84%)	247 (20.63%)
2-3 times/week	52 (12.18%)	179 (14.95%)
1-4 times/month	141 (33.02%)	365 (30.49%)
< 12 times/year	84 (19.67%)	260 (21.72%)
No alcohol last year	61 (14.29%)	146 (12.20%)
	Mean (95% CI)	*Mean (95% CI)*
Age (years)	46.7 (45.5, 47.9)	42.9 (42.3, 43.6)
Weight (kg)	72.6 (71.1, 74.2)	70.5 (69.6, 71.4)
Smoking intensity at baseline (cigarettes/day)	18.8 (17.6, 20.0)	21.2 (20.5, 21.8)
Years of smoking	26.6 (25.4, 27.8)	24.3 (23.6, 24.9)

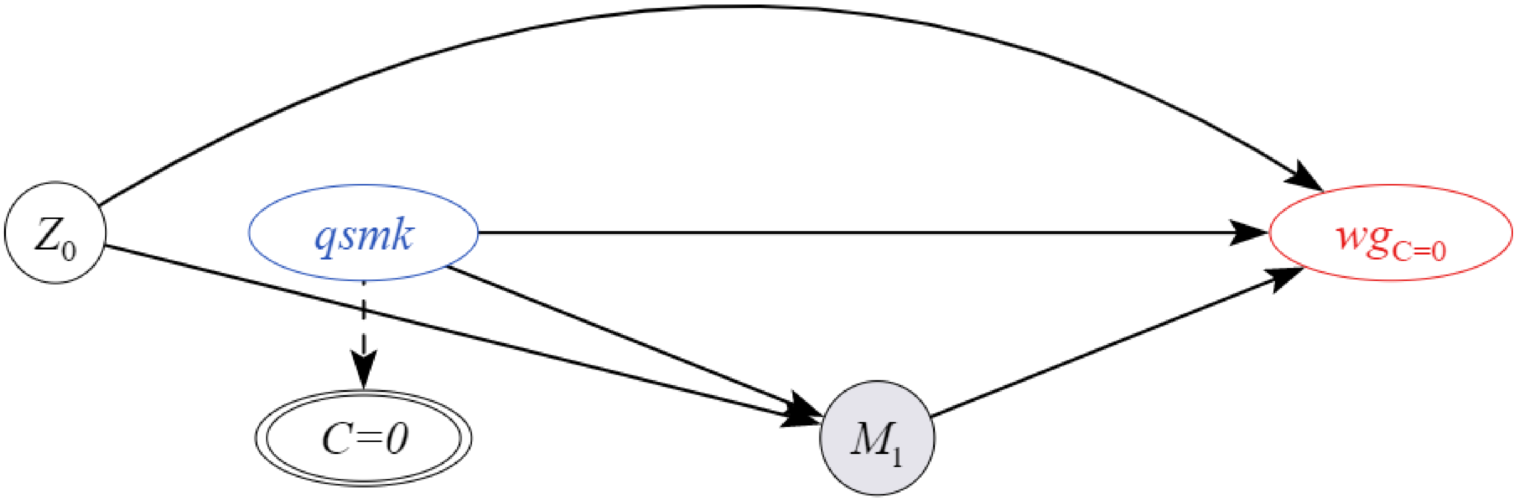

Figure 4.4: Simplified post-intervention directed acyclic graph (DAG) to depict the effect of using inverse probability of treatment weights with inverse probability of censoring weights to remove confounding and selection bias where Z_0 is the set of baseline confounders education, age, sex, race, recreational physical activity, daily life physical activity, years of smoking, smoking intensity, and weight at baseline. The outcome of interest is weight gain at follow-up among the uncensored. The joint exposure is smoking cessation status (q_{smk}) and the selection node *(C=0)* is included to depict censoring due to missing data in the outcome that is removed with inverse probability weighting. M_1 is the unmeasured set of post-smoking cessation mediators.

Table 4.3: Inverse-probability weighting for treatment and censoring adjusted estimates of the causal mean differences of weight gain on smoking cession status, NHEFS 1982-1984

	Model 1 (IPTW only)		***Model 2 (IPTW * IPCW)***	
	Estimate	*(Bootstrap 95% CI)*	*Estimate*	*(Bootstrap 95% CI)*
Quit smoking	3.52	(2.45, 4.49)	3.50	(2.43, 4.44)

*Abbreviations: IPTW = inverse probability treatment weighting; IPCW = inverse probability censoring weighting

Model 1 - IPTW-adjusted only for baseline confounders age, sex, race, education, smoking intensity, years of smoking, recreational physical activity, daily life physical activity, and weight

Model 2 - IPTW and IPCW-adjusted for baseline confounders age, sex, race, education, smoking intensity, years of smoking, recreational physical activity, daily life physical activity, weight, and smoking cessation to remove potential selection bias induced by censoring

Chapter 5 Discussion & References

5.1 Discussion

This dissertation aimed to guide the use of DAGs in settings relevant to applied health researchers. DAGs are considered to be causal diagrams and as such, they are taught from a causal perspective. However, even among epidemiologists, the consistent use of DAGs in applied health research remains low.[7,8] This may in part be due to a lack of clear guidance on the application of DAGs in non-causal settings such as studies with descriptive or predictive aims. All of these types of studies are commonly used in applied epidemiology but the understanding of how to apply DAGs in these types of settings may be low.

We provided steps and guidance on the use of DAGs in descriptive, predictive, and causal studies. To do so, we used one dataset from *Causal Inference: What If* to demonstrate the similarities and differences in the approach to answering these questions.[4,18] Though our examples were simplified we were able to focus on the relationship between smoking cessation and weight gain in the population to highlight the differences in approach and considerations for the three types of aims (**Table 5.1**). We showed that the estimates themselves are different and should not be conflated. We also found that for all three types of aims, the definition of the target population is critically important for inference. It can affect whether selection bias is an issue in any setting and should be corrected so the inference can be made on the target population. All studies have a descriptive component. Predictive aims can range from simply quantifying correlations to requiring causal assumptions to address the aim in full. Causal studies require that the conditions for identifiability be met. Using a DAG to depict existing knowledge can optimize data collection and variable selection. They clarify assumptions and can be augmented to provide valuable context that can aid in the identification of bias and the nature of missing data. They can also help guide inference from the estimate and communicate assumptions of the causal structure that may aid future research in assessing the transportability of the model/estimate to other target populations.

The applications we used to demonstrate the use of DAGs in descriptive, predictive, and causal aims were simple examples and did not touch on some potential issues that may arise such as identifying and correcting information bias, treating missing data, transporting the model to a new target population, effect modification and interaction, and approaching predictive models the various types of prognostic prediction questions with DAGs. The aim of this dissertation was to provide some guidance on the incorporation of DAGs to address some common aims in epidemiology and public health.

Applied epidemiologists can use DAGs to guide study design and analysis, optimize variable selection, and clarify assumptions. Krieger and Davey Smith have pointed out that being tied to a DAG can limit the scope of the study and may result in a failure to address true causes. However, in response, Pearl has pointed out that this in fact may be considered a strength of a DAG, to limit the scope of the study to the question of interest and later evolve to include new information and potential factors as they arise. In fact, a DAG can also illustrate the true relationship of the variables on the outcome. For example, race is a social construct that is often used in epidemiology to assess racial health disparities as a proxy for structural inequities. A DAG can better clarify the operationalization of race as a construct to better assess the true relationships between the societal mechanisms that make up race and the outcome.[103,104]

There remain limitations in how DAGs can be used or interpreted. DAGs are inherently limited to the assumptions that are presented in the DAG. Krieger and Davey Smith have pointed out that being tied to a DAG can limit the scope of the study and may result in a failure to address true causes.[105] However, in response, Pearl has pointed out that this in fact may be considered a strength of a DAG, to limit the scope of the study to the question of interest and later evolve to include new information and potential factors as they arise.[106] Similarly, they are limited to the question being asked when constructing the DAG and cannot completely account for potential confounders when evaluating the effects between other covariates in the graph with respect to the outcome. Attempting to extend a model that was constructed with or without a DAG beyond the exposure and outcome of interest can result in the Table 2 fallacy.[107] However, DAGs remain flexible tools that can be augmented and evolve to include necessary information over time.

5.2 Appendix

Table 5.1: Summary of application of descriptive, predictive, and causal aims

	Descriptive	***Predictive***	***Causal***
Research Question	Who is gaining weight? What is the crude difference in weight gain by smoking cessation status?	What are the strongest predictors of weight gain at follow-up	What is the average causal effect of smoking cessation on weight gain?
Outcome	Observed weight gain at follow-up	Predicted weight gain at follow-up	Weight gain at follow-up
Model	Simple linear regression of weight gain on smoking cessation status	Conditional inference random forest	Linear risk model with combined inverse probability of treatment weights and inverse probability of censoring weighting.
Target estimate	Crude mean and risk differences	Rank of important predictors of weight gain	Causal risk difference
Potential issues	Selection bias, information bias, missing data	Selection bias, information bias, missing data	Confounding, selection bias, information bias, missing data

5.3 References

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

2. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. *Epidemiology*. 1999;10(1):37-48.

3. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia; 2008.

4. Hernán MA, Robins JM. *Causal Inference: What If*. Chapman & Hall/CRC; 2020. Accessed July 6, 2022. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

5. Pearl J, Glymour M, Jewell NP. *Causal Inference in Statistics: A Primer*. John Wiley & Sons Ltd; 2016. Accessed July 6, 2022. https://www.wiley.com/en-us/Causal+Inference+in+Statistics%3A+A+Primer-p-9781119186847

6. Pearl J. An introduction to causal inference. *Int J Biostat*. 2010;6(2):Article 7. doi:10.2202/1557-4679.1203

7. Barnard-Mayers R, Childs E, Corlin L, et al. Assessing knowledge, attitudes, and practices towards causal directed acyclic graphs: a qualitative research project. *Eur J Epidemiol*. 2021;36(7):659-667. doi:10.1007/s10654-021-00771-3

8. Tennant PWG, Murray EJ, Arnold KF, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620-632. doi:10.1093/ije/dyaa213

9. Lash TL, Rothman KJ, VanderWeele TJ, Haneuse S. *Modern Epidemiology*. 4th ed. Wolters Kluwer; 2020.

10. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *J Clin Epidemiol*. 2022;142:264-267. doi:10.1016/j.jclinepi.2021.08.001

11. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29(4):722-729. doi:10.1093/ije/29.4.722

12. Ferguson KD, McCann M, Katikireddi SV, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. *Int J Epidemiol*. 2020;49(1):322-329. doi:10.1093/ije/dyz150

13. Lesko CR, Fox MP, Edwards JK. A framework for descriptive epidemiology. *Am J Epidemiol*. Published online July 1, 2022:kwac115. doi:10.1093/aje/kwac115

14. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing; 2019. doi:10.1007/978-3-030-16399-0

15. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

16. Petersen ML, van der Laan MJ. Causal Models and Learning from Data. *Epidemiology*. 2014;25(3):418-426. doi:10.1097/EDE.0000000000000078

17. National Center for Health Statistics. NHANES I - Epidemiologic Followup Study (NHEFS). Centers for Disease Control and Prevention. Published 2022. Accessed April 5, 2023. https://wwwn.cdc.gov/Nchs/Nhanes/Nhefs/Default.aspx

18. Hernán MA. Causal Inference: What If (the book). Miguel Hernan's Faculty Website. Published October 19, 2012. Accessed May 2, 2023. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

19. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. Basic Books; 2018.

20. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol*. 2020;35(7):619-630. doi:10.1007/s10654-020-00636-1

21. Ramspek CL, Steyerberg EW, Riley RD, et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol*. 2021;36(9):889-898. doi:10.1007/s10654-021-00794-w

22. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *Chance*. 2019;32(1):42-49. doi:10.1080/09332480.2019.1579578

23. Fox MP, Murray EJ, Lesko CR, Sealy-Jefferson S. On the Need to Revitalize Descriptive Epidemiology. *Am J Epidemiol*. 2022;191(7):1174-1179. doi:10.1093/aje/kwac056

24. Kelsey JL, Whittemore AS, Evans AS, Thompson WD. *Methods in Observational Epidemiology*. 2nd ed. Oxford University Press; 1996.

25. Koepsell TD, Weiss NS. *Epidemiologic Methods: Studying the Occurrence of Illness*. Oxford University Press; 2014.

26. Porta M, ed. *A Dictionary of Epidemiology*. 4th ed. Oxford University Press; 2014.

27. Conroy S, Murray EJ. Let the question determine the methods: descriptive epidemiology done right. *Br J Cancer*. 2020;123(9):1351-1352. doi:10.1038/s41416-020-1019-z

28. Platt RW. The importance of descriptive epidemiology. *Am J Epidemiol*. Published online August 25, 2022:kwac153. doi:10.1093/aje/kwac153

29. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(12):2705-2715. doi:10.1093/aje/kwy173

30. Westreich D. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press; 2019.

31. Peters J, Janzing D, Schölkopf B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press; 2017. Accessed November 8, 2022. https://library.oapen.org/handle/20.500.12657/26040

32. Piccininni M, Konigorski S, Rohmann JL, Kurth T. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC Med Res Methodol*. 2020;20(1):179. doi:10.1186/s12874-020-01058-z

33. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A*. 2016;113(27):7345-7352. doi:10.1073/pnas.1510507113

34. Pearl J, Bareinboim E. External Validity: From Do-Calculus to Transportability Across Populations. In: Geffner H, Dechter R, Halpern J, eds. *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Association for Computing Machinery; 2022:451-482. Accessed November 6, 2022. https://doi.org/10.1145/3501714.3501741

35. Steingrimsson JA, Gatsonis C, Li B, Dahabreh IJ. Transporting a Prediction Model for Use in a New Target Population. *American Journal of Epidemiology*. 2023;192(2):296-304. doi:10.1093/aje/kwac128

36. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15(3):413-419. doi:10.1093/ije/15.3.413

37. Robins JM, Greenland S. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*. 1992;3(2):143-155.

38. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov*. 2009;6:4. doi:10.1186/1742-5573-6-4

39. Glymour MM, Walter S, Tchetgen Tchetgen EJ. Natural experiments and instrumental variable analyses in social epidemiology. In: Oakes JM, Kaufman JS, eds. *Methods in Social Epidemiology*. 2nd ed. John Wiley & Sons; 2017.

40. Glymour MM, Swanson SA. Instrumental Variables and Quasi-Experimental Approaches. In: *Modern Epidemiology*. 4th ed. Lippincott Williams & Wilkins; 2020.

41. CDCTobaccoFree. Smoking & Tobacco Use: Fast Facts. Centers for Disease Control and Prevention. Published December 1, 2022. Accessed April 17, 2023. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm

42. National Library of Medicine. Smoking. MedlinePlus. Published January 25, 2019. Accessed April 17, 2023. https://medlineplus.gov/smoking.html

43. CDCTobaccoFree. Smoking Cessation: Fast Facts. Centers for Disease Control and Prevention. Published August 22, 2022. Accessed April 17, 2023. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/cessation/smoking-cessation-fast-facts/index.html

44. National Library of Medicine. Weight gain after quitting smoking: What to do. MedlinePlus. Published August 15, 2022. Accessed April 17, 2023. https://medlineplus.gov/ency/patientinstructions/000811.htm

45. Tips from Former Smokers. 7 Common Withdrawal Symptoms and What You Can Do About Them. Centers for Disease Control and Prevention. Published December 12, 2022. Accessed April 17, 2023. https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/7-common-withdrawal-symptoms/index.html

46. National Center for Health Statistics. *A Statistical Methodology for Analyzing Data from a Complex Survey: The First National Health and Nutrition Examination Survey*. U.S. Department of Health and Human Services; 1994. https://www.cdc.gov/nchs/data/series/sr_02/sr02_121.pdf

47. National Center for Health Statistics. *Statistical Issues in Analyzing the NHANES I Epidemiologic Followup Study*. U.S. Department of Health and Human Services; 1994. https://www.cdc.gov/nchs/data/series/sr_02/sr02_121.pdf

48. Schuessler J, Selb P. Graphical Causal Models for Survey Inference. Published online November 26, 2019. doi:10.31235/osf.io/hbg3m

49. National Center for Health Statistics. NHANES Tutorials - Weighting Module. Centers for Disease Control and Prevention. Published April 25, 2023. Accessed April 28, 2023. https://wwwn.cdc.gov/nchs/nhanes/tutorials/weighting.aspx

50. Keeney BJ, Fulton-Kehoe D, Wickizer TM, Turner JA, Chan KCG, Franklin GM. Clinically Significant Weight Gain One Year After Occupational Back Injury. *J Occup Environ Med*. 2013;55(3):318-324. doi:10.1097/JOM.0b013e31827943c6

51. Ball MP, Coons VB, Buchanan RW. A Program for Treating Olanzapine-Related Weight Gain. *PS*. 2001;52(7):967-969. doi:10.1176/appi.ps.52.7.967

52. Maina G, Albert U, Salvi V, Bogetto F. Weight gain during long-term treatment of obsessive-compulsive disorder: a prospective comparison between serotonin reuptake inhibitors. *J Clin Psychiatry*. 2004;65(10):1365-1371. doi:10.4088/jcp.v65n1011

53. Arah OA. Analyzing Selection Bias for Credible Causal Inference: When in Doubt, DAG It Out - PMC. *Epidemiology*. 2019;30(4):517-520. doi:10.1097/EDE.0000000000001033

54. Thompson CA, Arah OA. Selection bias modeling using observed data augmented with imputed record-level probabilities. *Ann Epidemiol*. 2014;24(10):747-753. doi:10.1016/j.annepidem.2014.07.014

55. Smith LH. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. *Curr Epidemiol Rep*. 2020;7(4):179-189. doi:10.1007/s40471-020-00241-6

56. SAS Institute Inc. SAS Software. Published online 2023.

57. Arah OA. Bias Analysis for Uncontrolled Confounding in the Health Sciences. *Annu Rev Public Health*. 2017;38:23-38. doi:10.1146/annurev-publhealth-032315-021644

58. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *Am J Epidemiol*. 2017;186(8):1010-1014. doi:10.1093/aje/kwx164

59. Thompson CA, Jin A, Luft HS, et al. Population-Based Registry Linkages to Improve Validity of Electronic Health Record-Based Cancer Research. *Cancer Epidemiol Biomarkers Prev*. 2020;29(4):796-806. doi:10.1158/1055-9965.EPI-19-0882

60. Arah OA, Chiba Y, Greenland S. Bias Formulas for External Adjustment and Sensitivity Analysis of Unmeasured Confounders. *Ann Epidemiol*. 2008;18(8):637-646. doi:10.1016/j.annepidem.2008.04.003

61. Rudolph KE, Stuart EA. Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods. *Am J Epidemiol*. 2018;187(3):604-613. doi:10.1093/aje/kwx248

62. Banack HR, Hayes-Larson E, Mayeda ER. Monte Carlo Simulation Approaches for Quantitative Bias Analysis: A Tutorial. *Epidemiol Rev*. 2021;43(1):106-117. doi:10.1093/epirev/mxab012

63. Greenland S, Pearl J. Causal Diagrams. In: *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd; 2017:1-10. doi:10.1002/9781118445112.stat03732.pub2

64. Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. *Eur J Epidemiol*. 2020;35(7):615-617. doi:10.1007/s10654-020-00659-8

65. VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology*. 2009;20(6):863-871. doi:10.1097/EDE.0b013e3181ba333c

66. Arah OA. Augmenting causal diagrams with effect modification, interaction and other parametric information. Presented at: Society for Epidemiologic Research - 48th Annual SER Meeting; June 2015; Denver, Colorado. Accessed February 15, 2023. https://epiresearch.org/wp-content/uploads/2015/07/Final-Abstract-Book.2.pdf

67. Attia J, Holliday E, Oldmeadow C. A proposal for capturing interaction and effect modification using DAGs. *Int J Epidemiol*. 2022;51(4):1047-1053. doi:10.1093/ije/dyac126

68. Gorelick MH. Bias arising from missing data in predictive models. *Journal of Clinical Epidemiology*. 2006;59(10):1115-1123. doi:10.1016/j.jclinepi.2004.11.029

69. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*. 2020;125:183-187. doi:10.1016/j.jclinepi.2020.03.028

70. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. doi:10.1136/bmj.b2393

71. Nijman S, Leeuwenberg A, Beekers I, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*. 2022;142:218-229. doi:10.1016/j.jclinepi.2021.11.023

72. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res*. 2012;21(3):243-256. doi:10.1177/0962280210394469

73. Mohan K, Pearl J. Graphical Models for Processing Missing Data. *JASA*. 2021;116(534):1023-1037.

74. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2023. Accessed May 19, 2023. https://www.r-project.org/

75. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28:1-26. doi:10.18637/jss.v028.i05

76. Kuhn M, Johnson K. Measuring Predictor Importance. In: Kuhn M, Johnson K, eds. *Applied Predictive Modeling*. Springer; 2013:463-485. doi:10.1007/978-1-4614-6849-3_18

77. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*. 2007;8(25). doi:https://doi.org/10.1186/1471-2105-8-25

78. Wirth KE, Tchetgen Tchetgen EJ. Accounting for selection bias in association studies with complex survey data. *Epidemiology*. 2014;25(3):444-453. doi:10.1097/EDE.0000000000000037

79. Balzer L, Petersen M, van der Laan M. Tutorial for Causal Inference. In: Bühlmann P, Drineas P, Kane M, Laan M van der, eds. *Handbook of Big Data*. CRC Press; 2016.

80. Matthay EC, Glymour MM. A Graphical Catalog of Threats to Validity: Linking Social Science with Epidemiology. *Epidemiology*. 2020;31(3):376-384. doi:10.1097/EDE.0000000000001161

81. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48(4):1294-1304. doi:10.1093/ije/dyz032

82. National Center for Health Statistics. NHANES I (1971-1974). Centers for Disease Control and Prevention. Accessed April 26, 2023. https://wwwn.cdc.gov/nchs/nhanes/nhanes1/default.aspx

83. Pearl J. Direct and indirect effects. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. ; 2022:373-392.

84. 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

85. Weinberg CR. Can DAGs clarify effect modification? *Epidemiology*. 2007;18(5):569-572. doi:10.1097/EDE.0b013e318126c11d

86. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology*. 2007;18(5):561-568. doi:10.1097/EDE.0b013e318127181b

87. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol*. 2007;166(9):1096-1104. doi:10.1093/aje/kwm179

88. Nilsson A, Bonander C, Strömberg U, Björk J. A directed acyclic graph for interactions. *Int J Epidemiol*. 2021;50(2):613-619. doi:10.1093/ije/dyaa211

89. Inoue K, Yan Q, Arah OA, et al. Air Pollution and Adverse Pregnancy and Birth Outcomes: Mediation Analysis Using Metabolomic Profiles. *Curr Environ Health Rep*. 2020;7(3):231-242. doi:10.1007/s40572-020-00284-3

90. Avin C, Shpitser I, Pearl J. *Identifiability of Path-Specific Effects*. UCLA; 2005. Accessed February 8, 2023. https://escholarship.org/uc/item/45x689gq

91. Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics*. 2014;71(1):1-14. doi:10.1111/biom.12248

92. Vansteelandt S, Daniel RM. Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology*. 2017;28(2):258-265. doi:10.1097/EDE.0000000000000596

93. Tai AS, Lin SH, Chu YC, Yu T, Puhan MA, VanderWeele T. Causal Mediation Analysis with Multiple Time-varying Mediators. *Epidemiology*. 2023;34(1):8-19. doi:10.1097/EDE.0000000000001555

94. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34(3):211-219. doi:10.1007/s10654-019-00494-6

95. Cinelli C, Forney A, Pearl J. A Crash Course in Good and Bad Controls. *Sociological Methods & Research*. Published online May 20, 2022:00491241221099552. doi:10.1177/00491241221099552

96. Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology*. 2011;22(5):745. doi:10.1097/EDE.0b013e318225c2be

97. DAGitty v3.0. Accessed May 9, 2023. http://www.dagitty.net/dags.html

98. Fusion. Accessed May 9, 2023. https://causalfusion.net/app

99. VanderWeele TJ, Arah OA. Unmeasured Confounding for General Outcomes, Treatments, and Confounders. *Epidemiology*. 2011;22(1):42-52. doi:10.1097/EDE.0b013e3181f74493

100. Helmich E, Boerebach BCM, Arah OA, Lingard L. Beyond limitations: Improving how we handle uncertainty in health professions education research. *Medical Teacher*. 2015;37(11):1043-1050. doi:10.3109/0142159X.2015.1073239

101. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *International Journal of Epidemiology*. 2020;49(1):338-347. doi:10.1093/ije/dyz251

102. Smith LH, Mathur MB, VanderWeele TJ. Multiple-bias Sensitivity Analysis Using Bounds. *Epidemiology*. 2021;32(5):625-634. doi:10.1097/EDE.0000000000001380

103. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. *The Annals of Family Medicine*. 2022;20(2):157-163. doi:10.1370/afm.2792

104. Howe CJ, Bailey ZD, Raifman JR, Jackson JW. Recommendations for Using Causal Diagrams to Study Racial Health Disparities. *Am J Epidemiol*. 2022;191(12):1981-1989. doi:10.1093/aje/kwac140

105. Krieger N, Davey Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol*. 2016;45(6):1787-1808. doi:10.1093/ije/dyw114

106. Pearl J. Comments on: The tale wagged by the DAG. *Int J Epidemiol*. 2018;47(3):1002-1004. doi:10.1093/ije/dyy068

107. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177(4). doi:10.1093/aje/kws412