Junior Management Science

www.jums.academy
ISSN: 2942-1861

Kausale Inferenz unter Anwendung von Double Machine Learning: Oregon Health Insurance Experiment

Sebastian Schmidt

University of Hamburg

Abstract

This study applies advanced methods of causal inference, specifically the Double Machine Learning (DML) framework, to estimate the causal effects of public health insurance on individual health outcomes using data from the 2008 Oregon Health Insurance Experiment (OHIE), a randomized controlled trial. DML integrates modern machine learning with the econometric principles of causal identification to obtain unbiased treatment effect estimates in high-dimensional data. Interactive Regression Models (IRM) and Interactive Instrumental Variable Models (IIVM) are employed to estimate effects of Medicaid coverage on perceived health, number of doctor visits, satisfaction, access to medical services, and quality of care. The results indicate small but positive causal effects of Medicaid coverage on perceived health and healthcare utilization, statistically insignificant effects on satisfaction and access to medication, and a slightly negative effect on perceived quality of care. The findings highlight the potential of Double Machine Learning as a robust framework for causal analysis in empirical research.

Zusammenfassung

Diese Arbeit verwendet fortgeschrittene Methoden der Kausalen Inferenz, insbesondere das Double Machine Learning (DML) Framework, um die kausalen Effekte f́6ffentlicher Krankenversicherung auf individuelle Gesundheitsvariablen anhand der Daten des Oregon Health Insurance Experiments (OHIE, 2008) zu sché4tzen. DML ermf́6glicht, unverzerrte Sché4tzungen von Treatment-Effekten in hochdimensionalen Datensé4tzen zu ermf́6glichen. Mittels Interactive Regression Models (IRM) und Interactive Instrumental Variable Models (IIVM) werden Effekte von Medicaid auf wahrgenommene Gesundheit, Anzahl der Arztbesuche, Zufriedenheit, Zugang zu medizinischen Leistungen sowie Versorgungsqualité4t untersucht. Die Ergebnisse zeigen geringe, aber positive kausale Effekte von Medicaid auf die wahrgenommene Gesundheit und die Hé4ufigkeit der Arztbesuche, statistisch insignifikante Effekte auf Zufriedenheit und Medikamentenzugang sowie einen leicht negativen Zusammenhang mit der wahrgenommenen Behandlungsqualité4t. Die Ergebnisse unterstreichen das Potenzial von Double Machine Learning als robustes Instrument der Kausalanalyse in empirischen Studien.

Keywords: causal inference; double machine learning; oregon health insurance experiment

1 Einleitung

In dieser Arbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mitgemeint, soweit es für die Aussage erforderlich ist. Englische Fachbegriffe ohne geläufiges deutsches Äquivalent werden im Original verwendet, da diese im Fachbereich gängig sind.

Eine der grundlegenden Fragen der Wissenschaft lautet „Warum?“. Seit Jahrtausenden treibt das Bestreben nach einer Erklärung für die beobachteten Phänomene der Welt Wissenschaftler sowie neugierige Denker an. Dabei kann diese Frage unzählig viele Facetten annehmen: Warum gewittert es? Warum fallen Äpfel auf den Boden? Warum werden Menschen krank? All das und viele weitere, sind Fragen nach der Ursächlichkeit bzw. der Kausalität. Es stellt sich die Frage, weshalb Kausalität so interessant für die Menschen ist. Und die Antwort liegt nahe: Das Verständnis von kausalen Zusammenhängen ermöglicht uns, Kausaleffekte zu reproduzieren, den Verlauf von Ereignissen in der Zukunft zu manipulieren und mit alledem zu unseren Gunsten zu wenden. Dieses Verständnis kann auf einer sehr oberflächlichen oder einer fundamentalen Ebene vorliegen - in beiden Fällen kann ein Mehrwert aus dem Erkennen kausaler Zusammenhänge geschlossen werden. „Das Essen dieser Pflanze führt zu Bauchschmerzen. Ich sollte diese Pflanze nicht mehr essen“ oder „Das Drehen einer metallischen Spule innerhalb eines magnetischen Felds induziert Elektrizität, die wir nutzen können, um elektrische Geräte mit Energie zu versorgen“. Das Wissen, dass das Eintreten von A anschließend das Eintreten von B auslöst oder begünstigt, bietet einen Mehrwert. Der Mensch neigt von Natur aus dazu, kausale Zusammenhänge schnell und intuitiv zu erlernen. Dennoch könnte es in

DOI: https://doi.org/10.5282/jums/v11i1pp107-138

© The Author(s) 2026. Published by Junior Management Science.
This is an Open Access article distributed under the terms of the CC-BY-4.0 (Attribution 4.0 International). Open Access funding provided by ZBW.

der Realität kaum schwieriger sein, kausale Zusammenhänge zu beweisen. Ist der Patient einer Studie tatsächlich genesen, weil er einer neuen Behandlungsmethode unterzogen wurde oder war es Zufall und der Patient wäre auch ohne medizinisches Eingreifen wieder gesund geworden?

Das akademische Feld, das sich damit beschäftigt Kausalität zu untersuchen, ist die Kausale Inferenz. Kausalität und damit auch Kausale Inferenz haben eine große Relevanz in vielen Bereichen des alltäglichen Lebens. Wie am zuvor genannten Beispiel ersichtlich, ist Kausalität enorm wichtig in der Medizin. Moderne Medizin baut generell darauf auf, dass der Behandlung einer Krankheit ein kausaler Effekt zur Genesung unterstellt wird - andernfalls gäbe es keinen Grund für die Anwendung einer Behandlung (der Placebo-Effekt ist dabei eingeschlossen, da auch dieser kausal ausgelöst werden kann und zur Genesung eines Patienten beiträgt) (Thompson & Upshur, 2017). Wäre Wissenschaftlern und Ärzten der volle kausale Zusammenhang zwischen sämtlichen Medikamenten, Behandlungen und Erkrankungen klar, so wären viele Krankheiten heilbar. Jedoch sind kausale Zusammenhänge häufig komplizierter als nur die Verabreichung eines Medikaments. Bei vielen Krankheiten kommt es beispielsweise auch darauf an, in welchem Stadium der Erkrankung eine Behandlung angewendet wird - in der Regel gilt hier, je früher die Therapie, desto höher die Chance auf Erfolg. Damit eine Krankheit rechtzeitig behandelt werden kann, muss sie auch rechtzeitig erkannt werden (Ott et al., 2009). Meist ist der Besuch eines Krankenhauses oder Arztes bei Auftreten erster Symptome dabei der entscheidende Faktor. Abhängig vom Gesundheitssystem des jeweiligen Landes, in dem potenzielle Patienten sich befinden, funktioniert der Besuch eines Arztes jedoch unterschiedlich. Einer der größten Unterschiede ist dabei die Abgrenzung zwischen privater und öffentlicher Gesundheitsversorgung. In Ländern wie Deutschland und Schweden sind die Einwohner standardmäßig über die staatliche öffentliche Krankenkasse versichert und können abgesehen von einem monatlichen Beitrag alle von der Krankenkasse anerkannten Leistungen kostenfrei oder unter geringer Kostenbeteiligung in Anspruch nehmen (§5 und §6 (*SGB V*, 1989)). Besuche bei niedergelassenen Ärzten und Krankenhäusern sind dabei gleichermaßen inbegriffen. In den USA hingegen ist das Gesundheitssystem weitgehend privatisiert. Dort gibt es kein öffentliches Gesundheitssystem, welches standardmäßig jeden Einwohner abdeckt. Der Default in den USA ist es, nicht krankenversichert zu sein und entweder eine private Krankenversicherung abzuschließen oder im Falle einer medizinischen Untersuchung oder Behandlung selbst für die Kosten aufzukommen (Vladeck, 2003). Eine Problematik bei diesem System ist, dass ein großer Anteil der Bevölkerung der USA sich den Abschluss einer privaten Krankenversicherung finanziell nicht leisten kann (Dickman et al., 2017). Generell sind medizinische Leistungen teuer - ein einfaches Röntgenbild kann tausende Dollar kosten, in Ausnahmefällen sogar mehr. Eine Chemotherapie zur Behandlung einer Krebserkrankung erreicht schnell den Wert eines Einfamilienhauses bis hin zu Millionenbeträgen (Kantarjian & Rajkumar, 2015). Das sind Kosten, die selbst wohlhabende Menschen schnell finanziell gefährden könnten. Dieser Umstand lässt Raum für Vermutung, dass die Art der Gesundheitsversicherung eine Reihe an Faktoren beeinflussen kann. Es wäre denkbar, dass Menschen ihr Verhalten daran anpassen, ob sie die Kosten für medizinische Leistungen selbst tragen müssen. Beispielsweise könnten die Anzahl an Arztbesuchen innerhalb eines Intervalls sowie der physische und psychische Gesundheitszustand der Bevölkerung und ähnliche Variablen von der Art der Versicherung beeinflusst sein. Der Vorteil von einem öffentlichen Gesundheitssystem ist, dass die Kosten für medizinische Behandlungen weitestgehend abgefangen werden. Darauf lässt sich die folgende Frage formulieren: „Welche direkten und indirekten kausalen Effekte hat Public Healthcare auf die abgedeckte Bevölkerung und wie stark sind diese?“. Bei dem Oregon Health Insurance Experiment (im Folgenden auch OHIE) aus dem Jahr 2008 wurden Daten zu genau der Beantwortung dieser Frage erhoben. Eine Auswahl an Personen aus dem Staat Oregon erhielten über einen Zeitraum von einem Jahr „MedicAid“ - also eine staatlich geförderte Gesundheitsversicherung, für die die Empfänger keine Kosten tragen mussten. Im Rahmen von mehreren Umfragen wurden im Verlauf des Experiments Daten zu den Teilnehmern erhoben. Die im OHIE gesammelten Daten wurden bereits mit traditionellen statistischen Verfahren analysiert (Baicker & Finkelstein, 2014).

Im Feld der Kausalen Inferenz sind innerhalb der letzten Jahre neue Ansätze und Methoden entwickelt worden, um Kausalität zu untersuchen. Diese Methoden sind vielversprechend, da sie einen neuen Blickwinkel auf die Daten aus dem OHIE und die assoziierten Fragen eröffnen. Eine dieser Methoden ist das Double Machine Learning Framework, welches versucht mithilfe von Machine Learning Algorithmen kausale Effekte zu berechnen (Chernozhukov et al., 2016). Seit Veröffentlichung des OHIE wurden bereits Studien und Untersuchungen mit den Daten aus dem Experiment erhoben, jedoch hat keine der wissenschaftlichen Arbeiten Double Machine Learning angewandt, um die potenziell zugrundeliegenden kausalen Effekte zu untersuchen. Im Rahmen dieser Arbeit soll genau das geschehen: Aus dem Blickwinkel der Kausalen Inferenz sollen die Daten des Oregon Health Insurance Experiment unter Verwendung vom Double Machine Learning Framework auf Kausalität untersucht werden. Dabei ist die Zielsetzung dieser Arbeit, die folgenden Leitfragen zu beantworten:

- Inwieweit hat MedicAid einen Einfluss auf die Gesundheit der Probanden?
- Ist die Anzahl der Arztbesuche innerhalb eines Zeitintervalls abhängig von MedicAid-Coverage?
- Beeinflusst MedicAid die Zufriedenheit der Probanden?

- Werden ohne MedicAid-Coverage eigentlich notwendige medizinische Leistungen nicht in Anspruch genommen?
- Hat MedicAid-Coverage einen Einfluss auf die Qualität der erhaltenen medizinischen Leistungen?

Die Arbeit ist dabei so strukturiert, dass zuerst die wissenschaftlichen Grundlagen und Definitionen dargelegt werden. Dafür wird im zweiten Kapitel der Fachbereich der Kausalen Inferenz näher erläutert. Begriffe wie Korrelation und Kausalität werden sinngemäß und mathematisch eingeführt. Zudem wird das Double Machine Learning Framework in seiner Notwendigkeit und seiner Funktionsweise erklärt. Im dritten Kapitel wird der in dieser Arbeit verwendete Datensatz des Oregon Health Insurance Experiments näher vorgestellt. Dabei wird zunächst die Durchführung des Experiments und anschließend der Datensatz beschrieben. Im vierten Kapitel wird die Methodik dieser Arbeit ausführlich aufgezeigt - zum einen wird dabei auf die konzeptionelle Vorgehensweise und zum anderen auf die technische Umsetzung eingegangen. Im fünften Kapitel werden die Ergebnisse der Anwendung des Double Machine Learning Frameworks ausgewertet. Mögliche Kritikpunkte an dem verwendeten Datensatz oder der angewandten Methodik werden im sechsten Kapitel diskutiert. Anschließend werden im siebten Kapitel die Ergebnisse zusammengefasst und es wird deren Validität bewertet. Im letzten Kapitel wird ein Ausblick für zukünftige Anwendung von Double Machine Learning - insbesondere im Kontext vom OHIE - gegeben.

2 Kausale Inferenz

Kausale Inferenz ist eine moderne Weiterentwicklung der Statistik. Jahrzehntelang wurden mithilfe von statistischen Methoden deskriptiv Wahrscheinlichkeiten, Zusammenhänge und Abhängigkeiten beobachtet und erforscht. Statistik selbst ist eine akademische Disziplin, die sich der Mathematik bedient und sich mit der Sammlung, Analyse, Interpretation, Präsentation und Organisation von Daten beschäftigt (*Oxford Dictionary of English*, 2010). Sie befasst sich damit, wie Daten gesammelt und so analysiert werden können, dass sich daraus sinnvolle Schlüsse ziehen lassen. Dabei geht es darum, Muster in den Daten zu identifizieren, Zusammenhänge zwischen Variablen aufzudecken und Hypothesen auf ihre Gültigkeit zu überprüfen (Romeijn, 2022). Doch ein entscheidender Punkt, der in der klassischen Statistik schwierig abzubilden ist, ist die Wirkungsrichtung von Ereignissen. Statistische Maße können beschreiben, dass zwei Ereignisse zusammenhängen - doch sie können nicht formulieren, inwieweit sie sich eventuell gegenseitig beeinflussen (Pearl, 2009a). Das Kapitel ist wie folgt strukturiert: Das erste Unterkapitel widmet sich den zum Verständnis dieser Arbeit notwendigen Definitionen. Im zweiten Unterkapitel wird das Double Machine Learning, ein Framework zur Anwendung von Machine Learning bei der Untersuchung kausaler Effekte, erklärt. Darin wird die Bedeutung des Frameworks für die Anwendung von Machine Learning sowie sein Aufbau erklärt.

2.1 Definitionen

Nachdem zuvor bereits die wissenschaftliche Disziplin der Kausalen Inferenz eingeführt wurde, ist es zunächst wichtig, die darin verwendeten Begrifflichkeiten zu definieren. Dabei sollten die Begriffe sowohl sinngemäß, als auch mathematisch abgegrenzt werden. Gerade Letzteres ist im Bereich der Statistik, Mathematik und der Kausalen Inferenz von großer Wichtigkeit. Um Berechnungen durchführen zu können, sollte zunächst ein Grundgerüst aus Definitionen und Formeln gebaut werden, um alle anschließenden Berechnungen zu stützen. In den folgenden Abschnitten soll es darum gehen, die relevanten Begrifflichkeiten der Kausalen Inferenz zu definieren und in Formeln abzubilden. Kernbegriff im Feld der Kausalen Inferenz ist natürlich die Kausalität. Diese soll - insbesondere in Abgrenzung zur Korrelation - definiert werden. Dafür geht es im ersten Unterkapitel um Korrelationen, deren Interpretation und eine Möglichkeit, diese abzubilden. Das zweite Unterkapitel grenzt daraufhin den Begriff der Kausalität ab und gibt jeweils die Formel für den individuellen und den durchschnittlichen kausalen Effekt. Zuletzt wird Confounding, eine potenzielle Störquelle bei der Bestimmung kausaler Effekte, erklärt und definiert.

2.1.1 Korrelation

Korrelationen sind in der Statistik von großer Bedeutung. Soll die Beziehung zwischen zwei oder mehr Variablen untersucht werden, ist die Korrelation meist das erste Werkzeug (Croxton & Cowden, 1939). Es gibt verschiedene Arten Korrelationen zu messen - abhängig von den betrachteten Variablen und der gestellten Frage können verschiedene Korrelationskoeffizienten verwendet werden. Der vermutlich gängigste ist der Pearson Korrelationskoeffizient. Dieser beschreibt die Stärke des linearen Zusammenhangs zwischen zwei Variablen. Der Koeffizient bewegt sich zwischen +1 und -1, wobei 1 eine perfekte positive Korrelation, -1 eine perfekte positive Korrelation und 0 keine Korrelation bedeutet. Der Pearson-Korrelationskoeffizient kann symbolisch dargestellt werden als:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

wobei r_{xy} der Pearson-Korrelationskoeffizient zwischen den beiden Variablen x und y ist, n die Stichprobengröße, x_i und y_i die Werte der beiden Variablen für jede Beobachtung i sind und $\bar{x}$ und $\bar{y}$ ihre jeweiligen Stichprobenmittelwerte sind (Wright, 1921). Während Korrelation nützlich ist, um die Beziehung zwischen zwei Variablen zu untersuchen ist jedoch wichtig zu beachten, dass eine Korrelation allein nicht zwangsläufig eine Kausalität zwischen den betrachteten Variablen impliziert (Wright, 1921). Während der Korrelation zwischen Rauchen und Lungenkrebs eine direkte Kausalbeziehung zugrundeliegt, ist dies bei der Korrelation

zwischen der Mortalität in Krankenhäusern und dem Volumen von Operationen anders. Im Jahr 2002 fanden Birkmeyer et al., dass die Mortalität für 14 Arten von Operationen invers mit der Häufigkeit der Operationen im Krankenhaus korreliert waren (Birkmeyer et al., 2002). Diese Korrelation ist jedoch unwahrscheinlich auf eine direkte kausale Beziehung zwischen der Häufigkeit der Operationen und der Mortalität der Patienten zurückzuführen. Stattdessen könnte sie auf andere Faktoren zurückzuführen sein, wie z.B. Erfahrung des Personals, die Ausstattung oder Ressourcenallokation des Krankenhauses. Wird eine Operation beispielsweise häufig durchgeführt, so sammeln die behandelnden Ärzte Erfahrung, welche sich positiv auf den Erfolg zukünftiger Eingriffe auswirken kann. Die Korrelation kann in diesem Fall also ein Indikator für Kausalität sein, jedoch stehen nicht alle korrelierten Variablen zwangsläufig in einer direkten Kausalbeziehung zueinander. Es besteht auch die Möglichkeit, dass Korrelationen lediglich zufällig Auftreten und gar keine Kausalität zugrunde liegt. Wie britischer Ökonom und Nobelpreisträger Ronald Coase bereits sagte: „If you torture the data long enough, it will confess to anything." (Coase, n. d.). Diese Aussage legt nahe, dass bei ausreichend langer und intensiver Datenanalyse fast jedes gewünschte Muster oder jede Verbindung gefunden werden kann, selbst wenn diese reinem Zufall geschuldet sind. Die Schlussfolgerung daraus ist, dass bei der Interpretation von statistischen Zusammenhängen stets versucht werden sollte, die Ergebnisse durch unabhängige Replikation und Validierung zu bestätigen (Roberts et al., 2006).

Neben der gerade erläuterten Problematik gibt es eine weitere Schwierigkeit beim Arbeiten mit Korrelationen: Korrelationskoeffizienten können keine Wirkungsrichtung abbilden (Pearl, 2009a). Um dies anhand eines Beispiels zu verdeutlichen, lässt sich der folgende Fall betrachten: Kopfschmerzen und die Einnahme von Schmerzmitteln sind positiv miteinander korreliert. In diesem Fall liegt auch eine kausale Beziehung vor. Ohne ein tieferes Verständnis für den Sachverhalt besteht jedoch keine Möglichkeit anhand des Korrelationskoeffizienten eine Aussage darüber zu treffen, welche der beiden Variablen die andere „auslöst". Mit mehr Wissen über den Sachverhalt oder einer Untersuchung der kausalen Koeffizienten wird ersichtlich, dass Kopfschmerzen begünstigen Schmerzmittel einzunehmen. Aufgrund der in diesem Abschnitt erläuterten Probleme bei der Verwendung von Korrelation zur Untersuchung kausaler Effekte, soll im folgenden Abschnitt Kausalität näher erläutert werden.

2.1.2 Kausalität und kausale Effekte

Die erste Lektion in jeder Statistik-Vorlesung und des vorigen Abschnitts ist die folgende: Korrelation bedeutet nicht gleich Kausalität. Kausalität bedeutet vereinfacht: Ein Ereignis beeinflusst oder verursacht ein anderes (Pearl, 2009a). Fällt ein Gärtner einen Baum, kippt dieser um. Das Bearbeiten des Baumes mit einer Axt oder Säge hat den Baum zum Umstürzen gebracht - somit war es der Auslöser. Es kann hier also von Kausalität gesprochen werden. Zumindest vermeintlich. Auch wenn es offensichtlich erscheint, dass der Baum lediglich umgefallen ist, weil der Gärtner ihn abgesägt hat, wäre es jedoch auch denkbar, dass der Baum im gleichen Moment ohne das Zutun des Gärtners auch umgefallen wäre. Stimmte dies, so wäre der Gärtner nicht der kausale Auslöser für das Fallen des Baumes. Kausalität liegt nämlich erst dann vor, wenn ein Ereignis A ein anderes Ereignis B auslöst und B ohne das Eintreten von A nicht passiert wäre (Pearl, 2009a). Das Ergebnis, auch bekannt als „Outcome", sollte abhängig davon unterschiedlich sein, ob ein Eingriff (in der Kausalen Inferenz „Treatment" genannt) durchgeführt wurde oder nicht. In der realen Welt stellt dies jedoch ein Problem dar: Entweder ein Ereignis tritt ein, oder nicht. Es kann entschieden werden, dass ein Treatment durchgeführt oder verabreicht wird - danach kann aber nicht mehr der Fall beobachtet werden, in dem das Treatment nicht verabreicht wurde. Wird ein Patient einer Operation unterzogen, so ist nach der Operation unmöglich zu erfahren, wie der Zustand desselben Patienten ohne den Eingriff gewesen wäre. Der in der Realität eingetretene Fall wird in der Kausalen Inferenz als „Factual" bezeichnet, während der gegensätzliche Fall das sogenannte „Counterfactual" ist (Pearl, 2009b). Es ist faktisch unmöglich beide Varianten zu beobachten. Ein Counterfactual ist per Definition nicht zu beobachten, was auch als das fundamentale Problem der Kausalen Inferenz bekannt ist (Holland, 1986). Obwohl Kausalität in der realen Welt nicht zu beweisen ist, kann sie dennoch untersucht werden. Dafür werden statistische Methoden zu Hilfe gezogen. Die bisher genannten Beispiele waren allesamt individuelle kausale Effekte - „Macht es einen Unterschied, ob wir diese Person mit einem Medikament behandeln, oder nicht?". In der Kausalen Inferenz werden durchschnittliche kausale Effekte untersucht - „Verabreichen wir dieses Medikament einer Gruppe von Erkrankten, haben diese dann eine höhere Chance auf Genesung, als die Kontrollgruppe?" (Pearl, 2009b).

Ist das Ziel kausale Effekte aus mathematischer Perspektive zu betrachten, so sind zunächst einige Festlegungen zu treffen. Diese wurden 1980 von Rubin unter dem Namen „Stable-Unit-Treatment-Value Assumption", im Folgenden als SUTVA abgekürzt, festgehalten (Granger, 1986; Rubin, 1980). SUTVA lässt sich im Wesentlichen auf zwei Aussagen reduzieren:

- Es darf keine Interferenzen zwischen den Probanden geben.
- Es darf keine verschiedenen Versionen des Treatments geben.

In beiden Fällen wäre Y_i^a nicht ausreichend definiert, um kausale Effekte zu bestimmen (Rubin, 1980). Erstere Aussage bedeutet, dass der Outcome des Treatments bei einem Probanden keinen Einfluss auf den Outcome eines anderen Probanden haben darf. Werden beispielsweise zwei Geschwister einer Herzoperation unterzogen, wäre es denkbar, dass die erfolgreiche Operation bei der ersten Person und die damit einhergehende Erleichterung bei Proband und Arzt einen positiven Effekt auf das Ergebnis der zweiten Operation hätten

(oder vice versa) und somit den kausalen Effekt verfälscht. Am selben Beispiel lässt sich auch die zweite Festlegung erklären. Es darf keine verschiedenen Versionen des Treatments geben: Das heißt, es muss für eine Vergleichbarkeit die exakt selbe Operation an allen Probanden durchgeführt werden. Wird bei Proband *A* ein Herzschrittmacher eingesetzt und bei Proband *B* wird ein Bypass gelegt, so sind die Eingriffe nicht vergleichbar miteinander. Patienten denen ein Herzschrittmacher eingesetzt wurde, sind jedoch untereinander vergleichbar. In der tiefsten Bedeutung dieser Annahme steckt, dass Treatments zu 100% übereinstimmen müssten, bis ins kleinste Detail. Unterscheidet sich der behandelnde Arzt, müsste von verschiedenen Treatments gesprochen werden. Streng genommen würde sich das Treatment bereits unterscheiden, wenn eine Operation 5 Sekunden länger dauert als eine andere oder der behandelnde Arzt bereits einen langen Arbeitstag hinter sich hat (und bei einer anderen Operation nicht). Es mag dem praktischen Nutzen nicht dienen, die Definitionen so streng zu befolgen, aber jegliche Abmilderung der Annahmen sind potenziell Fehlerquellen bei der Ermittlung von Kausalität (Rubin, 1980). In der Regel ist trotz bester Intentionen eine so strenge Durchführung jedoch nicht möglich und es wird versucht, Treatments so vergleichbar wie realistischerweise möglich zu halten. Im Nachhinein kann dann bewertet werden, ob von einem konstanten Treatment gesprochen werden kann bzw. ob die Variationen im Treatment irrelevant für die Bestimmung des kausalen Effekts sind. Im Fall des Oregon Health Insurance Experiments sind die SUTVA-Annahmen jedoch weitgehend erfüllt. Das in dieser Arbeit betrachtete Treatment ist eine Aufnahme in das MedicAid-Programm. Das MedicAid-Programm und die dazugehörigen Leistungen sind für alle Teilnehmer identisch und die Aufnahme in das Programm hat keinen Einfluss darauf, ob eine andere Person ebenfalls aufgenommen wird oder nicht.

Nachdem die wesentlichen Annahmen getroffen sind, kann begonnen werden, kausale Effekte durch mathematische Formeln zu definieren: Es gibt eine dichotome bzw. binäre Treatment Variable, im Folgenden *A* genannt; $A = 1$ bedeutet, das Treatment wurde angewandt, $A = 0$ steht für unbehandelt. Das Wort „Treatment“ kann dabei auch auch außerhalb des medizinischen Kontextes verstanden werden. Das Ausstrahlen eines Werbespots kann ebenfalls ein Treatment sein. Der Outcome wird in der Variable *Y* festgehalten, welche in diesem Fall ebenfalls dichotom ist. Geht es um die Wirksamkeit eines neuen Medikaments, so ergäbe sich daraus beispielsweise folgende Übersicht:

A = 0: Das Medikament wird nicht verabreicht
A = 1: Das Medikament wird verabreicht
Y = 0: Der Patient wird nicht geheilt
Y = 1: Der Patient wird geheilt

Daraus lässt sich die Formel für den individuellen kausalen Effekt wie folgt formulieren:

Das Treatment *A* hat einen kausalen Effekt auf den Outcome *Y*, wenn

$$Y^{a=1} \neq Y^{a=0} \tag{2}$$

(Hernan & Robins, 2023)

Wie zuvor bereits erwähnt, ist lediglich eine der beiden Optionen, das Factual, in der Realität beobachtbar - entweder das Treatment wird verabreicht, oder nicht. Die Formel besagt, dass Factual und Counterfactual verschieden voneinander sein müssen, um zu beweisen, dass ein kausaler Effekt vorliegt - per Definition ist dies jedoch im individuellen Fall nicht messbar. Kausale Effekte können dennoch untersucht werden, indem der durchschnittliche kausale Effekt betrachtet wird. Der durchschnittliche kausale Effekt ist der Effekt, welcher in der Grundgesamtheit zu beobachten wäre. Mathematisch lässt sich der durchschnittliche kausale Effekt wie folgt ausdrücken:

Ein durchschnittlicher kausaler Effekt liegt vor, wenn

$$E(Y^{a=1}) \neq E(Y^{a=0}) \tag{3}$$

(Hernan & Robins, 2023)

Das *E()* steht dabei für den statistischen Erwartungswert. Die Formel besagt also, dass ein durchschnittlicher kausaler Effekt vorliegt, wenn der Erwartungswert des Outcomes bei Verabreichung des Treatments sich von dem Erwartungswert des Outcomes bei Nicht-Verabreichung des Treatments unterscheidet. Interpretieren lässt sich das für das obige Beispiel als: „Wenn die durchschnittliche Erwartung einer Genesung durch die Anwendung des Treatments anders ist als ohne das Treatment, dann lässt sich von einem kausalen Effekt sprechen“. In der Realität wird allerdings selten die Grundgesamtheit untersucht, da dies häufig einen großen Aufwand darstellt oder teilweise schlichtweg nicht möglich ist. Es wird mit Stichproben gearbeitet (Chambers, 2003). Es ist jedoch zu berücksichtigen, dass die Untersuchung von Stichproben eine Fehlerquelle für Zufallsfehler darstellt. Um valide Schlussfolgerungen ziehen zu können, ist es daher wichtig, sich dieser Beschränkung bewusst zu sein und die Stichprobe repräsentativ für die Grundgesamtheit zu wählen (Chambers, 2003).

Eine weitere Quelle für Zufallsfehler sind nicht-deterministische Factuals bzw. Counterfactuals (Hernan & Robins, 2023). Nicht-deterministisch bedeutet in diesem Kontext, dass das Ergebnis eines Treatments nicht vorbestimmt ist. Liegen $Y^{a=1}$ und $Y^{a=0}$ beispielsweise eine stochastische Komponente zugrunde, hat ein Treatment nur eine Wahrscheinlichkeit zu wirken. Wird in 100 hypothetischen Fällen dieselbe Person *X* einem Treatment unterzogen, könnte das Treatment in nur 80 dieser Fälle wirksam sein - obwohl alle anderen Faktoren identisch gehalten sind. Das Treatment hat in diesem Fall also eine hohe Wahrscheinlichkeit zu wirken, aber es ist nicht garantiert. Nicht-deterministische Treatments führen demnach eine Unsicherheit in die Betrachtung ein und erschweren dadurch die isolierte Betrachtung kausaler Effekte. Es wurde

vermutet, dass Kausalzusammenhänge in der Realität nicht-deterministisch sein können (Yarlett, 2002). So lässt sich beispielsweise beim regelmäßigen Rauchen von Zigaretten kein deterministischer Zusammenhang zur Entstehung von Lungenkrebs feststellen, dennoch besteht ein Anhaltspunkt für einen kausalen Zusammenhang (Davey Smith, 2009), (Resnik & Vorhaus, 2006). Die zwei zuvor erläuterten Quellen für Zufallsfehler lassen sich zum Abschluss dieses Kapitels wie folgt festhalten:

- Das Verwenden von Stichproben bedeutet, dass nicht $E(Y^{a=1})$ und $E(Y^{a=0})$ bzw. $P(Y^{a=1})$ und $P(Y^{a=0})$ beobachtet werden können, sondern

$$\tilde{P}(Y^{a=1}) \text{ und } \tilde{P}(Y^{a=0}) \tag{4}$$

- Die Möglichkeit nicht-deterministischer Outcomes unter Bedingung von $a=0$ bzw. $a=1$ besteht.

(Hernan & Robins, 2023)

2.1.3 Confounding

In der Kausalen Inferenz geht es darum, den kausalen Effekt, den ein Treatment D potenziell auf einen Outcome Y hat, zu ermitteln. In der Realität spielen jedoch häufig viele Faktoren zusammen und es ist nicht offensichtlich, ob ein Outcome Y ausschließlich von dem Treatment D beeinflusst wurde (Hernan & Robins, 2023). Gibt es eine Variable, welche sowohl das Treatment, als auch den Outcome beeinflusst, so wird von einem Confounder gesprochen (VanderWeele & Shpitser, 2013). Wird beispielsweise in einer Beobachtungsstudie, also einer Studie in welcher die Daten nicht aus einem randomisierten Experiment stammen, der Effekt von Rauchen auf Herzkrankheiten untersucht, sind mögliche Confounder denkbar.

Im folgenden Beispiel ist Rauchen das Treatment, Übergewicht ein Confounder und die Herzkrankheit der Outcome. Angenommen Übergewicht hat einen negativen Effekt auf die Herzgesundheit, bzw. erhöht die Wahrscheinlichkeit eine Herzerkrankung zu erleiden. Beeinflusst Übergewicht jetzt zusätzlich noch das Rauchen, so liegt ein Confounder vor. Übergewicht könnte ein Indikator für die Vernachlässigung der eigenen Gesundheit sein. Menschen, die ihre Gesundheit vernachlässigen, sind möglicherweise weniger abgeschreckt von den negativen Konsequenzen des Rauchens. In diesem fiktiven Beispiel wäre die Wahrscheinlichkeit, dass übergewichtige Menschen zudem auch rauchen höher, als bei nicht übergewichtigen Leuten. In diesem Fall wirkt der Confounder „Übergewicht“ über zwei Pfade auf Y. Zum einen die direkte Auswirkung, die Übergewicht auf den Outcome hat, weil das Risiko für Herzkrankheiten dadurch erhöht wird und zum anderen die indirekte Auswirkung - da Übergewicht die Wahrscheinlichkeit erhöht zu rauchen, erhöht es zudem indirekt auch die Wahrscheinlichkeit eine Herzerkrankung zu erleiden.

Im Falle von Confounding können kausale Effekte nicht präzise bestimmt werden (Greenland et al., 1999b). Es ist unklar, wie viel des Einflusses, der auf Y wirkt, auf den kausalen Effekt des Treatments D und wie viel davon auf den Confounder X zurückzuführen ist. Die Verbindung zwischen Treatment und Outcome über den Confounder wird Backdoor-Path genannt. Bleibt ein Backdoor-Path unerkannt oder unberücksichtigt, kann dies zu einer Verzerrung des geschätzten kausalen Effekts führen (Greenland et al., 1999a). Um an dieser Stelle dennoch einen kausalen Effekt des Treatments bestimmen zu können, muss auf den Confounder X, in diesem Fall das Übergewicht, bedingt werden (Greenland et al., 1999a). „Bedingen“ (oder auch „kontrollieren“) bedeutet in der Statistik, eine Variable konstant zu halten, sodass keine Ungenauigkeit durch Änderungen in der Variable mehr entstehen kann (Durrett, 2019). Im vorigen Szenario könnten lediglich Personen, die einen BMI von Z haben, in die Betrachtung eingehen. Da der BMI und damit auch der Faktor „Übergewicht“ nun bei allen Probanden identisch ist, wurde der Confounding-Bias behoben.

2.2 Double Machine Learning

Nachdem zuvor die theoretische Grundlage im Bereich der Kausalen Inferenz gelegt wurde, soll nun das im Rahmen dieser Arbeit angewandte Double Machine Learning Framework erklärt werden. Machine Learning generell ist ein relevantes Werkzeug in Zeiten von Big Data (Wang & Alexander, 2016). Riesige Datenmengen sind heutzutage in vielen Bereichen zum Standard geworden, was bei klassischen statistischen Methoden, wie beispielsweise der linearen Regression, zu Problemen führen kann. Während die lineare Regression zwar von mehr Beobachtungen in den Daten profitiert, neigt die lineare Regression gerade bei hochdimensionalen Daten (d.h. ein hohes Verhältnis von Variablen zu Beobachtungen) zum Overfitting, also dem „Auswendiglernen“ der Datenpunkte (Hawkins, 2003). Andere Machine Learning Modelle profitieren hingegen häufig deutlich von großen Datenmengen, da sie darauf ausgelegt sind, komplexe, nicht-lineare Beziehungen zwischen Variablen zu erlernen, ohne starke Annahmen über die zugrundeliegende Verteilung der Daten zu treffen. Zudem haben viele Machine Learning Algorithmen Mechanismen, mit denen Overfitting entgegengewirkt werden kann (Zhou et al., 2017).

Datensätze, welche für die Untersuchung von kausalen Effekten verwendet werden, sind in der Regel ebenfalls umfangreich, da alle im Kausalzusammenhang stehenden Variablen in den Datensatz aufgenommen werden müssten (Hernan & Robins, 2023). Deshalb ist es naheliegend, dass auch in diesem Bereich Machine Learning einen Einzug findet. Es gibt grundlegend zwei Ansätze Machine Learning für die Schätzung kausaler Effekte anzuwenden: „Naive Machine Learning“ und „Double Machine Learning“. Naive Machine Learning bedeutet, dass ein Machine Learning Modell verwendet wird, um anhand der Kovariablen X den Effekt von dem Treatment D auf den Outcome Y zu schätzen. Dabei entsteht

potenziell der sogenannte Regularization Bias. Ein Bias ist eine systematische Abweichung bzw. Verzerrung der Ergebnisse vom wahren Sachverhalt (Georgii, 2009). Double Machine Learning hingegen ist ein Framework, das darauf abzielt, Machine Learning Algorithmen für Kausale Inferenz zu verwenden, ohne dabei einen Bias zu erzeugen (Chernozhukov et al., 2018). In den folgenden Abschnitten wird erst die Problematik von Naive Machine Learning - konkret, der Regularization Bias - erläutert. Im Anschluss daran wird Double Machine Learning aufbauend auf dem vorigen Abschnitt erklärt.

2.2.1 Naive Machine Learning und Regularization Bias

Wie bereits erwähnt, ist der Regularization Bias Grund dafür, dass Naive Machine Learning verzerrte Ergebnisse bei der Schätzung kausaler Effekte liefern kann. So kann das Auslassen von Confoundern dazu führen, dass das verwendete Modell den zugrundeliegenden Kausalzusammenhang nicht ausreichend abbildet (Hernan & Robins, 2023). Regularization ist eine gängige Methode, um die Komplexität der Machine Learning Algorithmen zu „bestrafen“ und dadurch Overfitting bzw. die Varianz des Modells zu reduzieren (Tibshirani, 1996). Die Problematik daran ist, dass dadurch ein Bias in die Ergebnisse eingeführt wird. Die Einführung eines Bias bedeutet in diesem Kontext, dass der zugrundeliegende Sachverhalt nun mit einem Modell abgebildet wird, das simpler ist, als der eigentliche Sachverhalt. Da das Modell potenziell zu simpel für den Sachverhalt ist, kann es diesen nicht akkurat approximieren - das wird auch Underfitting genannt. Sowohl Overfitting, als auch Underfitting (durch Regularization) führen zu verzerrten Modellen bei der Anwendung von Naive Machine Learning (Chernozhukov et al., 2018). Es gilt, den Kompromiss zwischen Bias und Varianz zu finden, bei dem der Generalisierungsfehler im Modell am geringsten ist - dieser Zusammenhang ist auch als Bias-Varianz-Tradeoff bekannt (Kohavi & Wolpert, 1997). Der durch Anwendung von Naive Machine Learning entstehende Regularization Bias kann auch durch Simulationsdaten visualisiert werden. Es gilt folgendes Teil-lineares Modell:

$$\begin{aligned} y_i &= \theta_0 d_i + g_0(x_i) + \zeta_i, \quad \zeta_i \sim \mathcal{N}(0,1), \\ d_i &= m_0(x_i) + v_i, \qquad\qquad v_i \sim \mathcal{N}(0,1). \end{aligned} \tag{5}$$

Es soll der kausale Koeffizient θ_0 geschätzt werden. Dabei sind die Kovariablen durch $x_i \sim \mathcal{N}(0, \Sigma)$ gegeben und $m_0(x_i)$ und $g_0(x_i)$ sind die Störfunktionen. Da es sich um Simulationsdaten handelt, können die Parameter Σ und θ_0 vom Anwender festgelegt werden.

In der Abbildung 1 wurde Naive Machine Learning auf den simulierten Datensatz angewandt. Dabei steht $\hat{\theta}$ für die Schätzung des kausalen Effekts, θ für den reellen kausalen Effekt und $\hat{\sigma}$ für die Standardabweichung. Die Gleichung $(\hat{\theta} - \theta)/\hat{\sigma}$ beschreibt also die standardisierte Abweichung des geschätzten kausalen Effekts von dem wirklichen kausalen Effekt. Da in diesem Fall der reelle kausale Effekt bekannt

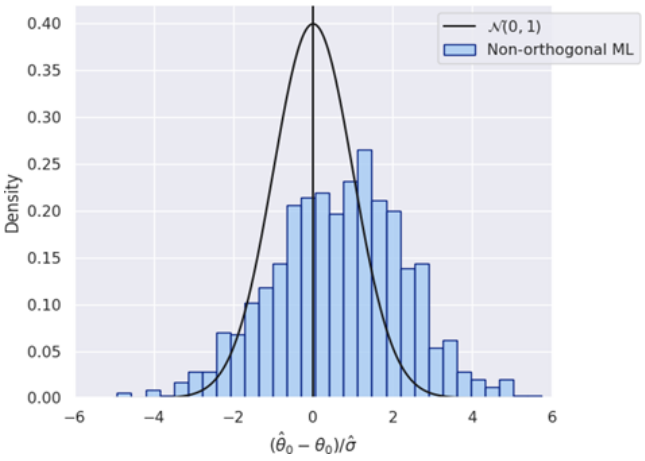

Abbildung 1: Regularization Bias durch Anwendung von Naive Machine Learning (Quelle: (Bach et al., 2022))

ist, kann so die Fähigkeit von Modellen und Frameworks evaluiert werden, diesen zu schätzen. Liegt in dem Modell kein Bias vor, wäre zu erwarten, dass das Histogramm wie der Zufallsfehler verteilt ist (in diesem Fall normalverteilt) - in der Abbildung ist jedoch deutlich eine Abweichung von dieser Normalverteilung zu sehen. Die Anwendung von Naive Machine Learning riskiert einen Bias einzuführen, welcher die resultierenden Schätzungen der kausalen Effekte verzerrt (Chernozhukov et al., 2018).

2.2.2 Double/Debiased Machine Learning

In diesem Abschnitt soll das Double Machine Learning Framework erläutert werden. Double Machine Learning (gelegentlich auch Debiased Machine Learning) ist eine mehrstufige Methode, um robuste Schätzungen für kausale Effekte zu erzeugen (Chernozhukov et al., 2016). Die folgende Gleichung beschreibt das Modell:

$$Y_i = d_i \alpha_0 + x_i' \beta + \epsilon \tag{6}$$

- Y_i ist der Outcome
- d_i ist das Treatment
- z_i sind mögliche Confounder
- x_i ist die Liste aller Kontrollvariablen (schließt z_i und Interaktionen ein)
- α ist der Koeffizient von d_i bzw. der kausale Effekt, den es zu schätzen gilt
- ϵ ist ein Störterm

Zunächst wird der Datensatz in k verschiedene Sample unterteilt - das nennt sich Crossfitting oder auch Sample Splitting und dient dazu, Overfitting zu vermeiden. Auf jedem der Sample werden zwei Machine Learning Modelle trainiert. Das erste Modell schätzt die Treatment-Variable D, das zweite Modell die Outcome-Variable Y. In beiden

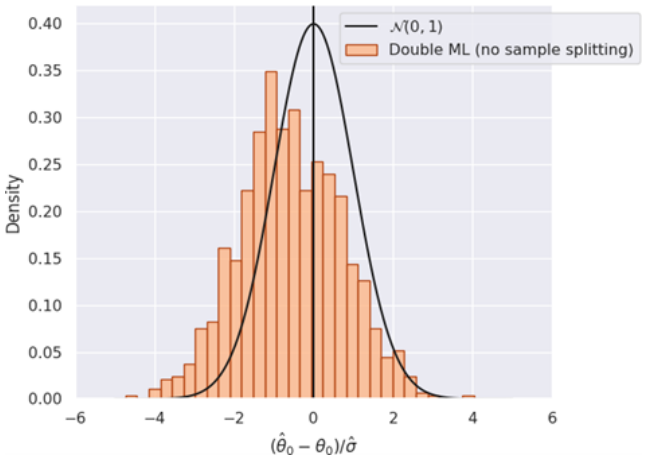

Abbildung 2: Double Machine Learning - ohne Crossfitting (Quelle: (Bach et al., 2022))

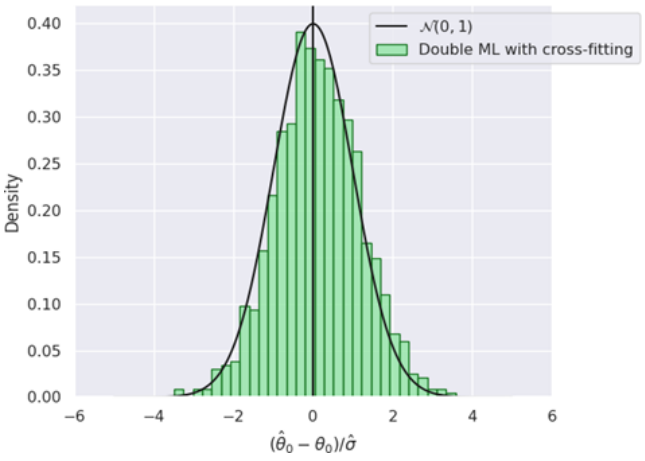

Abbildung 3: Double Machine Learning - mit Crossfitting (Quelle: (Bach et al., 2022))

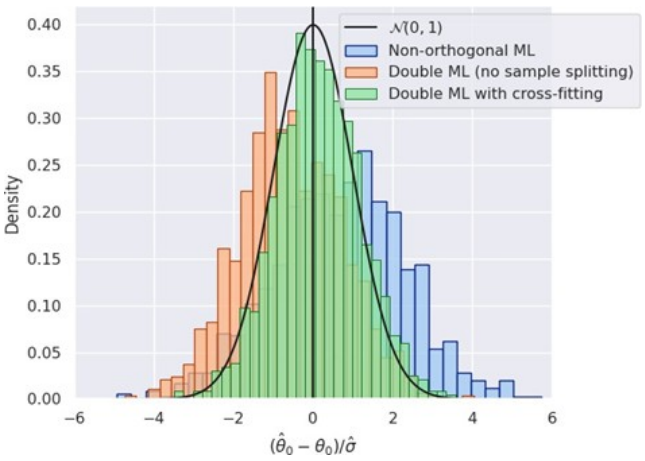

Abbildung 4: Vergleich: Naive ML, Double ML, Double ML Crossfitting (Quelle: (Bach et al., n. d.))

Fällen werden die Schätzungen auf der Grundlage der Kontrollvariablen X getroffen. Die Schätzungen werden als $\hat{D}_k$ und $\hat{Y}_k$ respektive gekennzeichnet. Anschließend werden die sogenannten Residuen $r_{D_k} = D_k - \hat{D}k$ und $r_{Y_k} = Y_k - \hat{Y}_k$, also die Differenzen zwischen den geschätzten und den eigentlichen Werten, errechnet. Nachdem die Residuen für jedes Sample errechnet wurden, wird ein weiteres Verfahren (in der Regel OLS) angewandt - dabei werden die Residuen der Treatment-Variable r_{D_k} als die unabhängige und die Outcome-Residuen r_{Y_k} als die abhängige Variable verwendet. Der von OLS geschätzte Koeffizient α_k ist der kausale Effekt, den das Treatment (im Sample) auf den Outcome hat. Die Ergebnisse aller k Samples werden zusammengeführt, um eine robustere Schätzung des kausalen Effekts α zu erhalten. Zusätzlich werden noch Hypothesentests angewandt, um die statistische Signifikanz der Ergebnisse zu bewerten. Sind die Ergebnisse signifikant, so kann von einem kausalen Effekt des Treatments auf den Outcome gesprochen werden (Chernozhukov et al., 2016, 2018).

In der Abbildung 2 ist zunächst einmal die Performance von Double Machine Learning ohne die Verwendung von Crossfitting dargestellt. Es ist erkennbar, dass die Qualität der Schätzungen sich im Vergleich zum naiven Ansatz verbessert hat - der Bias ist kleiner. Jedoch ist nach wie vor eine Verzerrung der Ergebnisse ersichtlich. Dieser Bias entsteht in diesem Fall durch Overfitting.

Um den Bias zu beheben, kann Crossfitting angewandt werden. Die Ergebnisse für Double Machine Learning unter der Verwendung von Crossfitting können in der Abbildung 3 gesehen werden.

Bei dem Vergleich der drei Verfahren in der Abbildung 4 ist ersichtlich, dass Double Machine Learning unter Verwendung von Sample Splitting die besten Ergebnisse liefert. Diese Erkenntnis ist konsistent mit den Ergebnissen von Chernozhukov et al. in ihrer Publikation „Double/debiased machine learning for treatment and structural parameters“ (Chernozhukov et al., 2018). Im Rahmen dieser Arbeit wird ebenfalls mit Double Machine Learning unter Verwendung von Crossfitting gearbeitet.

3 Oregon Health Insurance Experiment

Das Oregon Health Insurance Experiment ist eine amerikanische Studie, bei der die Auswirkungen von einer Ausweitung öffentlicher Krankenversicherung auf eine weitreichende Auswahl an Variablen, wie zum Beispiel Gesundheit, Wohlbefinden und finanzielle Lage der Teilnehmer, untersucht wurde. Die Studie wurde im Jahr 2008 in Oregon als randomisiertes Experiment mit Kontrollgruppe durchgeführt. Die erhobenen Daten und die Ergebnisse der Studie sind öffentlich frei verfügbar und wurden bereits in Studien näher untersucht. Ziel der wissenschaftlichen Arbeiten ist es, mehr über die Vor- und Nachteile der Ausweitung öffentlicher Krankenversicherung in den Vereinigten Staaten zu sammeln (Baicker & Finkelstein, 2014). In den kommenden Abschnitten wird das Experiment und die damit verbundenen Daten genauer betrachtet, um eine

zielgerichtete und nachvollziehbare Analyse der kausalen Effekte durchzuführen. Für die korrekte Anwendung Kausaler Inferenz ist es von hoher Bedeutung, den untersuchten Datensatz und seine Zusammenhänge zu verstehen. Daher werden im ersten Unterkapitel der Hintergrund und das Lotterieverfahren des OHIE beschrieben. Im zweiten Abschnitt folgt eine Erläuterung der verschiedenen Datensätze des Experiments und einer Selektion an relevanten Variablen.

3.1 Lotterie

Der goldene Standard des wissenschaftlichen Arbeitens ist ein randomisiertes Experiment (Chalmers et al., 1981). Eine nicht-zufällige Auswahl der Probanden einer Studie könnte potenziell einen Bias in die Ergebnisse einführen - beispielsweise durch den sogenannten Selection Bias (gelegentlich auch Sampling Bias). Dieser tritt auf, wenn die untersuchte Teilgruppe die Grundgesamtheit nicht akkurat widerspiegelt. Haben bestimmte Personengruppen eine höhere Wahrscheinlichkeit in das Sample einzugehen als andere, so verzerrt dies das Verhältnis in Relation zur Grundgesamtheit und verzerrt somit auch die Ergebnisse (Heckman, 1977). Das Erheben von zufälligen Stichproben gewährleistet in der Theorie ein Sample frei von Bias, weshalb Ziel in der Konzeption einer jeden Studie sein sollte, die Probanden so randomisiert wie möglich auszuwählen (Chalmers et al., 1981).

Beim Oregon Health Insurance Experiment wurden die Probanden durch ein Lotterieverfahren ausgewählt, um eine zufällige Auswahl zu gewährleisten. Die Studie umfasste etwa 90.000 Personen, die auf der Warteliste eines ansonsten geschlossenen MedicAid-Programms standen, von denen ca. 30.000 Menschen in insgesamt acht Verlosungen ausgewählt wurden. Die ausgewählten Personen hatten die Möglichkeit, sich für eine öffentliche Krankenversicherung namens Oregon Health Plan Standard zu bewerben. Oregon Health Plan Standard richtet sich dabei an einkommensschwache Erwachsene im Alter von 19-64 Jahre, welche keinen Anspruch auf andere MedicAid-Programme oder ähnliche öffentliche Krankenversicherungen in Oregon haben. Um an der Lotterie teilnehmen zu können, mussten die Kandidaten bestimmte Kriterien erfüllen, wie z.B. einen Wohnsitz in Oregon, ein Einkommen unterhalb der Armutsgrenze und Ersparnisse unter 2.000 Dollar, sowie ein andauernder Zeitraum von mindestens sechs Monaten, währenddessen der Kandidat nicht krankenversichert war. Zusätzlich zur eigenen Bewerbung bestand auch die Möglichkeit weitere Haushaltsmitglieder in die Bewerbung einzubeziehen (Baicker & Finkelstein, 2014).

Zunächst wurde die Liste der Bewerber von Duplikaten bereinigt. Anschließend wurden Bewerber, die nicht den festgelegten Kriterien entsprachen, aussortiert. Dazu gehörten Personen, die vor Ablauf der Studie das maximale Alter von 64 Jahren erreicht hätten, Personen, die nicht in Oregon wohnhaft waren und Bewerber, die vor Beginn der Studie bereits verstorben waren. Nach der Bereinigung des Datensatzes blieben noch 74.922 zulässige Bewerber übrig, unter denen Plätze im Oregon Health Plan Standard verlost wurden. Insgesamt haben 29.834 Menschen in der Lotterie gewonnen und erhielten somit die Möglichkeit, sich für das MedicAid-Programm zu bewerben. Die Möglichkeit auf Bewerbung für das MedicAid-Programm wurde zusätzlich auch auf jedes Haushaltsmitglied der Gewinner ausgeweitet. Wenn zum Beispiel eine Person aus einem Zweipersonenhaushalt in der Lotterie gezogen wurde, konnten sowohl der Gewinner als auch sein Lebenspartner oder Mitbewohner sich für MedicAid bewerben, selbst wenn Letzterer nicht an der Verlosung teilgenommen hatte (Baicker & Finkelstein, 2014). Streng genommen werden dadurch die Annahmen des randomisierten Experiments verletzt, da nicht mehr alle Teilgruppen eine identische Wahrscheinlichkeit haben, im Datensatz zu erscheinen (Heckman, 1977). Hat beispielsweise jede Person in einem vierköpfigen Haushalt separat an der Verlosung teilgenommen, so hätte es gereicht, wenn eine Person als Gewinner gezogen wird, damit der gesamte Haushalt sich für das MedicAid-Programm bewerben könnte. Demnach hätte jede Person des Haushalts eine viermal so hohe Chance, in die Versuchsgruppe einzugehen, wie eine einzeln teilnehmende Person. Unter Beachtung dieser Tatsache ist die Versuchsgruppe des Oregon Health Insurance Experiments nur dann zufällig verteilt, wenn auf die Anzahl der Mitglieder eines Haushalts, welche auf der Warteliste stehen, bedingt wird. Es ist also relevant, bei der Auswertung des Datensatzes auf die angemeldeten Haushaltsmitglieder zu bedingen, um einen Sampling Bias zu vermeiden. Da diese Variable jedoch im Datensatz erfasst wurde, ist es möglich diese bei der Auswertung der Ergebnisse einzubeziehen. Da dieser Umstand in dieser Arbeit berücksichtigt und kontrolliert wurde, ist jedoch davon auszugehen, dass kein Bias durch den Versuchsaufbau in die Ergebnisse eingegangen ist.

3.2 Überblick über die Daten

In diesem Abschnitt wird der Fokus auf die bei den Teilnehmern gemessenen Variablen gelegt, welche im Rahmen mehrerer Umfragen erhoben wurden. Wenn mehr Variablen bei der Untersuchung kausaler Effekte erfasst werden, ist es wahrscheinlicher, alle relevanten Variablen für den Kausalzusammenhang zu berücksichtigen (Hernan & Robins, 2023). Der Datensatz des OHIE ist so groß, dass im Rahmen dieses Kapitels nicht jede Variable separat erklärt werden kann. Stattdessen wird eine Übersicht über eine Auswahl an potenziell relevanten Variablen gegeben, um ein allgemeines Verständnis für die später folgenden Auswertungen zu gewährleisten. Es ist jedoch möglich, dass nicht-erläuterte Variablen Teil späterer Analysen werden. Um einen umfassenden Überblick über die Variablen des OHIE und ihre Bedeutung zu erhalten, wird empfohlen, den Userguide der Studie zu konsultieren. In diesem finden sich ebenfalls deskriptive Analysen jeder einzelnen Variable. Die Daten des OHIE sind in einem relationalen Datenbankformat in mehrere Datensätze unterteilt. Jeder Teilnehmer wurde mit einer einzigartigen Identifikationsnummer versehen, die als Primärschlüssel in allen

Datensätzen verwendet wird. Diese ist in der Variable **person_id** gespeichert.

Der Datensatz *oregonhie_descriptive_vars.dta* enthält demographische Informationen der Teilnehmer sowie Angaben darüber, ob sie in der Lotterie ausgewählt wurden und ihren Bewerbungsstatus. Wichtige Variablen aus dem Datensatz sind die **person_id**, die **household_id** und das **treatment**. Letzteres ist eine Binärvariable, die Auskunft darüber gibt, ob ein Teilnehmer in der Lotterie gewonnen hat. Drei Variablen des Datensatzes geben verschiedene Zeitpunkte, die im Zusammenhang mit der Lotterie stehen an: **dt_notify_lottery** ist das Datum, an dem Gewinner der Lotterie darüber benachrichtigt wurden, dass sie sich für das MedicAid Programm bewerben können. **dt_app_decision** dokumentiert das Datum, an dem über die Bewilligung der Bewerbung eines Antragstellers entschieden wurde. Da der Versicherungsschutz über das MedicAid-Programm rückwirkend ausgestellt wurde, beschreibt **dt_retro_coverage** das Datum, ab dem der Versicherungsschutz einer Person aktiv wurde - da die gesamte Lotterie in mehreren Etappen ausgeführt wurde, kann sich das Datum in Abhängigkeit davon, welche Iteration der Lotterie gewonnen wurde, unterscheiden. Für die Gewinner der ersten Auslosung am 10. März 2008 wurde der Versicherungsschutz beispielsweise rückwirkend für alle medizinischen Kosten aktiviert, die am oder nach dem 11. März 2008 entstanden sind, obwohl die eigentlichen Bewerbungen der Teilnehme erst zu einem späteren Zeitpunkt bewilligt wurden. Wenn ein Teilnehmer im Studienzeitraum, also nach Genehmigung seiner Bewerbung und vor dem 1. Januar 2010, verstorben ist, wird dies in der Variable **postn_death** vermerkt. Dabei werden jedoch nur Todesfälle innerhalb von Oregon erfasst. Es ist theoretisch denkbar, dass verstorbene Personen, die außerhalb von Oregon gestorben sind, nicht gekennzeichnet sind. (Baicker & Finkelstein, 2014).

Bei *oregonhie_survey0m_vars.dta* handelt es sich um den Datensatz, welcher die Ergebnisse einer ersten Umfrage beinhaltet. Die erste von drei schriftlichen Umfragen wurde zwischen Juni und November im Jahr 2008 durchgeführt. Eine Stichprobe von 58.405 Personen bekam die Umfrage zugeschickt - davon waren 29.589 Menschen Teil der Versuchsgruppe, die anderen 28.816 gehörten zur Kontrollgruppe. Insgesamt haben 26.423 Teilnehmer auf die Umfrage geantwortet. Die nächste der Umfragen wurde sechs Monate nach der initialen Befragung durchgeführt - der dazugehörige Datensatz heißt *oregonhie_survey6m_vars.dta*. Die Durchführung sowie die abgefragten Variablen sind identisch mit denen aus der ersten Umfrage. 6.359 Teilnehmer antworteten auf diese Umfrage (Baicker & Finkelstein, 2014). Die letzte der drei Umfragen, deren Ergebnisse im Datensatz
oregonhie_survey12m_vars.dta gespeichert sind, wurde 12 Monate nach der initialen Befragung durchgeführt. Bei den Umfragen wurde ein monetärer Anreiz gesetzt, um die Teilnahmequote zu erhöhen - 5 Dollar bei Abschluss der Umfrage und zusätzlich die Chance 200 Dollar bei einer Verlosung zu gewinnen. Da die Antwortquote mit 36% recht gering war, wurde noch ein Nachbearbeitungsprotokoll gestartet - von den Personen, welche nicht geantwortet haben, wurden 30% ausgewählt und mehrmals telefonisch sowie postalisch kontaktiert und mit weiteren monetären Anreizen angesprochen. Dadurch konnte die Zahl der eingegangenen Antworten auf insgesamt 23.777 angehoben werden, was einer Antwortquote von 41% bzw. einer gewichteten Antwortquote von 50% entspricht (Baicker & Finkelstein, 2014).

Es gibt zwei weitere Datensätze, welche speziell für weitere Studien angefertigt wurden. Beide Datensätze werden im Verlauf dieser Arbeit nicht untersucht, weshalb sie nicht detaillierter vorgestellt werden. Im Rahmen dieser Arbeit wird der Datensatz
oregonhie_survey12m_vars.dta untersucht, da dieser größer ist, als als der Datensatz *oregonhie_survey6m_vars.dta*. Der in dieser Arbeit bereitgestellte Python-Code zur Bereinigung und Aufbereitung der Daten ist jedoch universell für alle Datensätze des OHIE anwendbar. Mit geringen Anpassungen lassen sich sämtliche Untersuchungen der dritten Umfrage auch auf die anderen Datensätze übertragen.

4 Methodik

In den vorigen Kapiteln wurden sowohl die Grundlagen der Kausalen Inferenz, als auch das Oregon Health Insurance Experiment erläutert. In diesem Kapitel soll das Vorgehen zur der Untersuchung kausaler Effekte in dieser Arbeit beschrieben werden. Das vorliegende Kapitel hat den Zweck, dem Leser dabei zu helfen, die verschiedenen Entscheidungsprozesse nachzuvollziehen, die bei der Erstellung dieser Arbeit getroffen wurden. Insbesondere wird dabei berücksichtigt, dass die Wissenschaft der Kausalen Inferenz sowie das Teilgebiet des Double Machine Learnings noch sehr jung sind. Das Kapitel soll somit zukünftigen Arbeiten als Orientierungshilfe dienen. Im ersten Unterkapitel wird das technische Grundgerüst der Arbeit, insbesondere die Wahl der Programmiersprache und die verwendeten Packages, beleuchtet. Im zweiten Unterkapitel werden die verwendeten Modelle und der Versuchsaufbau näher beschrieben. Zur Bestimmung kausaler Effekte sind verschiedene Algorithmen und Verfahren anwendbar und es hängt von dem Datensatz sowie den dahinterliegenden Modellannahmen ab, welche davon die besten Ergebnisse liefern. Dementsprechend wird darauf eingegangen, welche Erwägungen und Annahmen im Rahmen dieser Arbeit getroffen und gegeneinander getestet wurden. Im dritten Unterkapitel wird das gesamte Preprocessing der Daten in Python detailliert dargestellt. Das Preprocessing ist wichtig, damit die Daten in einer Form vorliegen, die von den Machine Learning Modellen verarbeitet werden kann. Im vierten Unterkapitel wird RandomForest, der Machine Learning Algorithmus der in dieser Arbeit fürs Double Machine Learning verwendet wird, vorgestellt. Schließlich wird im fünften und

letzten Unterkapitel Hyperparameter-Tuning erklärt und aufgezeigt, in welchem Umfang dies im Rahmen dieser Arbeit verwendet wurde.

4.1 Python

In dieser Arbeit wird Python als Programmiersprache verwendet, um mögliche kausale Effekte im Oregon Health Insurance Experiment zu untersuchen. Die erfolgreiche Installation von Python und aller verwendeten Bibliotheken ist für die Durchführung und Reproduktion der Untersuchung notwendig. Python bietet eine Reihe an Vorteilen, die die Programmiersprache besonders geeignet für den Umgang mit Daten machen. Einer dieser Vorteile ist die Vielfalt an Open-Source-Modulen, auch Packages genannt, die mit minimalem Aufwand importiert werden können. Ein Modul ist in der Regel eine Sammlung von Funktionen, die zur Lösung thematisch verwandter Anwendungsfälle eingesetzt werden können. Die wichtigsten Packages, die in dieser Arbeit verwendet werden, sind *pandas*, *DoubleML* und *sklearn*. *pandas* ist ein Modul, das eine Vielzahl von Funktionen zur Datenverarbeitung bereitstellt. Es ermöglicht sowohl das Auslesen zahlreicher Dateiformate als auch die Umwandlung in das eigene DataFrame-Format von *pandas*. *pandas* wird hier verwendet, um die *.dta-Formate* aus dem OHIE zu lesen und die Weiterverarbeitung zu erleichtern. Das *DoubleML*-Package ist eine Implementierung des Double Machine Learning Frameworks von Chernozhukov et al. und ist sowohl für R als auch für Python verfügbar (Chernozhukov et al., 2016). Das Modul basiert auf dem bekannten Package *sklearn*, das häufig im Bereich Deep Learning und Machine Learning eingesetzt wird. Die Verwendung von *DoubleML* erleichtert die Untersuchung kausaler Effekte, da bereits implementierte Verfahren und Modelle direkt genutzt werden können. Andere verwendete Module oder Python-Syntax werden im Verlauf der Arbeit nur oberflächlich erklärt, um eine allgemeine Nachvollziehbarkeit zu gewährleisten. Für ein tieferes Verständnis wird auf die Dokumentationen der entsprechenden Packages verwiesen.

4.2 Modelle und Versuchsaufbau

Grundlegend gibt es in der Kausalen Inferenz vier verschiedene Modelle, welche potenziell implementiert werden können (Bach et al., n. d.):

- Partially Linear Regression Model (PLR)
- Partially Linear IV Regression Model (PLIV)
- Interactive Regression Model (IRM)
- Interactive IV Model (IIVM)

Welches davon für den vorliegenden Anwendungsfall geeignet ist, kann im Wesentlichen anhand von zwei Unterscheidungen getroffen werden. Die Unterscheidung besteht zum einen darin, ob das Treatment binär verteilt ist und zum anderen darin, ob Instrumentalvariablen im Modell sind, beziehungsweise vermutet werden. Im Falle eines nicht-binären Treatments ohne Instrumentalvariablen würde ein Partially Linear Regression Model verwendet werden. Da das Treatment in dem Oregon Health Insurance Experiment binär verteilt ist, lässt sich dieses Modell für die Anwendung auf den Datensatz ausschließen - es muss ein Interactive Regression Model oder ein Interactive IV Model verwendet werden. Welches der beiden Modelle bestenfalls angewandt werden sollte, hängt dabei davon ab, ob Instrumentalvariablen vorliegen (Chernozhukov et al., 2017). Instrumentalvariablen werden in der Wissenschaft verwendet, um in einem nicht-randomisierten Experiment kausale Effekte schätzen zu können. Hat eine Variable einen direkten Effekt auf eine andere Variable, kann sie als Instrumentalvariable verwendet werden - in der Regel bedeutet dies, das Instrument und Treatment eine hohe Korrelation aufweisen (Angrist & Imbens, 1995). Es ist jedoch wichtig, dass die Instrumentalvariable nicht mit dem Fehlerterm korreliert ist, da dies sonst zu einer Verzerrung der Ergebnisse führen würde (Nichols, 2006).

Liegen in dem Modell keine Instrumentalvariablen vor, so wird das Interactive Regression Model verwendet. Wird hingegen von Instrumentalvariablen ausgegangen, so sollte das Interactive IV Model verwendet werden. Im Fall des Oregon Health Experiments ist nicht eindeutig zu sagen, welches Modell besser geeignet ist. In der Theorie handelt es sich bei dem OHIE um ein randomisiertes Experiment, was die Verwendung von Instrumentalvariablen überflüssig macht. Auf der anderen Seite ist der Versuchsaufbau des OHIE nur dann ein randomisiertes Experiment, wenn auf die Anzahl an Mitgliedern pro Haushalt der Teilnehmer bedingt wird (Baicker & Finkelstein, 2014). Außerdem ist nicht ausgeschlossen, dass es bei der Datenerhebung anderweitig Fehlerquellen gab. Aus diesen Gründen werden im Rahmen dieser Arbeit beide Modelle angewandt und verglichen.

Der Aufbau ist dabei wie folgt:

A. Interactive Regression Model

Interactive Regression Models haben folgende Form

$$\begin{aligned} Y &= g_0(D,X) + U, \quad \mathbb{E}(U \mid X, D) &&= 0, \\ D &= m_0(X) + V, \quad \mathbb{E}(V \mid X) &&= 0 \end{aligned} \tag{7}$$

wobei das Treatment $D \in \{0,1\}$ binär verteilt ist (Bach et al., 2021). Die visualisierte Darstellung der Kausalbeziehung kann in Abbildung 5 gesehen werden. Im Oregon Health Insurance Experiment bezeichnet die Treatment-Variable, ob eine Person im Rahmen der Lotterie ausgewählt wurde. Wurde eine Person ausgewählt, so erhielt sie die Chance, sich auf MedicAid zu bewerben. Ob letztendlich tatsächlich eine erfolgreiche Bewerbung eingereicht wurde, ist in der Treatment-Variable jedoch nicht festgehalten. Ziel dieser Arbeit ist zu untersuchen, ob Public Healthcare einen kausalen Effekt auf ausgewählte Variablen hat. Demnach wird die Tatsache, dass ein Teilnehmer MedicAid erhalten hat, als kausales Treatment gesetzt. Dafür wird die Variable **approved_app** verwendet (genaueres im Kapitel 4.3). Die

ursprüngliche Treatment-Variable des OHIE wird verwendet, um die sogenannte „Intention to treat“ zu messen. Die Intention to treat misst den Effekt, den die Absicht das Treatment einzuführen, hat. Zusätzlich dazu wird das IRM ebenfalls verwendet, um den eigentlichen Treatment-Effekt zu schätzen.

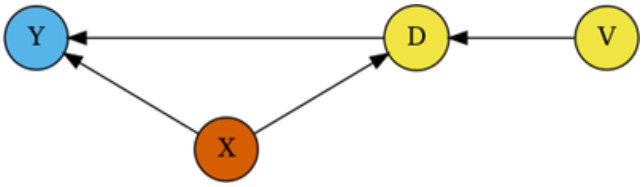

Abbildung 5. Kausal-Diagramm IRM (Quelle: (Bach et al., 2021))

B. Interactive IV Model

Interactive IV Models haben folgende Form

$$\begin{aligned} Y &= \ell_0(D,X) + \zeta, & \mathbb{E}(\zeta \mid Z,X) &= 0, \\ Z &= m_0(X) + V, & \mathbb{E}(V \mid X) &= 0 \end{aligned} \tag{8}$$

wobei das Treatment $D \in \{0,1\}$ und die Instrumentalvariablen $Z \in \{0,1\}$ binär verteilt sind (Bach et al., 2021). Die visualisierte Darstellung der Kausalbeziehung kann der Abbildung 6 entnommen werden. Mit dem IIVM wird ausschließlich der Treatment-Effekt geschätzt. Die Auswahl in der Lotterie wird dabei als Instrumentalvariable für den Erhalt von MedicAid verwendet.

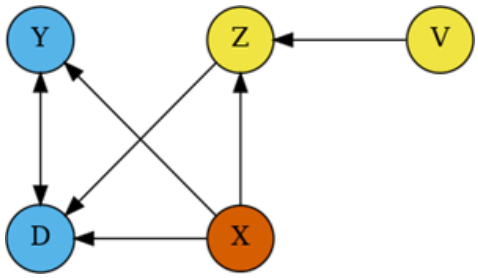

Abbildung 6. Kausal-Diagramm IIVM (Quelle: (Bach et al., 2021))

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert).

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert).

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert). Da das IIVM theoretisch mit Daten aus Beobachtungsstudien arbeiten kann, wird zusätzlich der vollständige Datensatz ohne Kontrolle auf **numhh** verwendet. Im Folgenden werden die Datensätze als „randomisierter/bedingter Datensatz“ und „vollständiger Datensatz“ betitelt.

Die Intention to treat wird in dieser Arbeit als Baseline für den kausalen Effekt von Public Healthcare verwendet - der gemessene kausale Effekt für das Treatment sollte sich also von der Baseline unterscheiden, damit davon gesprochen werden kann, dass das Treatment tatsächlich kausale Auswirkungen auf die untersuchten abhängigen Variablen hat. Sind der Treatment-Effekt und der Effekt der Intention to treat identisch, so hat das Treatment keinen kausalen Effekt auf die betrachteten abhängigen Variablen. Um sowohl für das IRM als auch das IIVM optimale Bedingungen zu gewähren, wird im Rahmen dieser Arbeit mit zwei verschieden verarbeiteten Datensätzen gearbeitet. Wie zuvor geschildert wurde, muss der Datensatz des OHIE auf die Variable **numhh** bedingt werden, um ein randomisiertes Experiment zu gewährleisten. Der erste Datensatz entsteht also daraus, dass die gesamte Datenmenge auf die Variable **numhh** bedingt wird (d.h. es wird auf eine einzelne Ausprägung der Variable gefiltert). Da das IIVM theoretisch mit Daten aus Beobachtungsstudien arbeiten kann, wird zusätzlich der vollständige Datensatz ohne Kontrolle auf **numhh** verwendet. Im Folgenden werden die Datensätze als „randomisierter/bedingter Datensatz“ und „vollständiger Datensatz“ betitelt.

Für die Implementation der Modelle wird das *DoubleML*-Package verwendet. Zusätzlich zur Wahl der kausalen Modelle können Machine Learning Algorithmen spezifiziert werden. In der Praxis kann es abhängig von der Modellwahl zu leichten Abweichungen bei den Schätzungen des kau-

salen Effekts kommen. Im Rahmen dieser Arbeit wird der RandomForest-Algorithmus verwendet, weil dieser in der wissenschaftlichen Literatur vielfach erprobt ist und meist sehr robuste Ergebnisse liefert - auch ohne Hyperparameter-Tuning (Probst et al., 2019). „Robust“ bedeutet im diesem Kontext, dass der Algorithmus trotz Ausreißern, fehlenden Werten oder anderen Arten von Fehlern oder Anomalien in den Daten seine Genauigkeit beibehalten kann. Im gleichnamigen Unterkapitel werden der Algorithmus und seine Hyperparameter genauer erklärt (Syrgkanis & Zampetakis, 2020).

4.3 Datenverarbeitung

In diesem Kapitel wird der gesamte Prozess vom Import der Daten, Import der Python Libraries und die Bereinigung der Daten veranschaulicht. Alle verwendeten Packages sind in dem verwendeten Python-Environment bereits zuvor installiert werden und können deshalb mit einem *Import*-Statement abgerufen werden. Für die Replikation dieses Versuchs müsste demnach eine Python-Umgebung angelegt werden.

Zunächst werden alle benötigten Packages importiert. Wie in einem vorigen Kapitel bereits erwähnt, ermöglicht *pandas* die Handhabung und Umformung von Daten. Die *DoubleML*-Library wird für das Double Machine Learning verwendet. Das Package *sklearn* enthält den RandomForest-Algorithmus.

```
import pandas as pd
from doubleml import DoubleMLData
from doubleml import DoubleMLIIVM, DoubleMLIRM
from sklearn.ensemble import RandomForestRegressor,
    RandomForestClassifier
from sklearn.base import clone
```

Im nächsten Schritt werden die Datensätze des Oregon Health Insurance Experiments importiert und zusammengeführt. Diese liegen im Stata- bzw. *.dta*-Format vor und können mit der Funktion *pandas.read_stata()* in das *pandas.DataFrame*-Format überführt werden. Importiert wird der Datensatz mit den deskriptiven Variablen zu den Teilnehmern und die drei durchgeführten Umfragen. Anschließend werden die Datensätze zu den Umfragen jeweils mit dem Datensatz der deskriptiven Variablen zusammengeführt.

```
# import datasets as pandas.DataFrame
d_vars = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_descriptive_vars.dta")
s_0m = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_survey0m_vars.dta")
s_6m = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_survey6m_vars.dta")
s_12m = pd.read_stata("oregon_data/OHIE_Data/
    oregonhie_survey12m_vars.dta")

df_0 = pd.merge(s_0m, d_vars, how='inner', left_on='
    person_id', right_on='person_id')
df_6 = pd.merge(s_6m, d_vars, how='inner', left_on='
    person_id', right_on='person_id')
df_12 = pd.merge(s_12m, d_vars, how='inner', left_on='
    person_id', right_on='person_id')
```

Anschließend gilt es, die Datensätze in eine Form zu bringen, mit der weitergearbeitet werden kann. In der Regel bedeutet dies, dass nicht-numerische Variablen in numerische Form umgewandelt werden. Zuvor sollte jedoch auf Missing Values geprüft werden. In Python werden fehlende Werte als *NaN* ("not a number") gekennzeichnet. Es gibt verschiedene Möglichkeiten mit Missing Values zu verfahren:

- Zeilen mit Missing Values löschen
- Spalten mit Missing Values löschen
- Missing Values auffüllen
- Missing Values ignorieren

Jede der Optionen hat prinzipiell verschiedene Szenarien in denen sie mehr oder weniger sinnvoll ist. Da im Oregon Health Insurance Experiment eine große Menge an *NaN* vorliegt, ist es voraussichtlich nicht sinnvoll die Werte lediglich aufzufüllen (beispielsweise mit dem Modus/Mittelwert), da dies mitunter stark die Ergebnisse verzerren könnte (Acuna & Rodriguez, 2004). Reihen bzw. Beobachtungen zu eliminieren ist ebenfalls dann sinnvoll, wenn die Quote an *NaN* hoch genug ist oder wenn die jeweilige Person bei entscheidenden Variablen keinen Wert hat - fehlt beispielsweise der Wert in der untersuchten abhängigen Variable, kann die Beobachtung nicht in die Auswertung eingehen (Acuna & Rodriguez, 2004). An dieser Stelle ist jedoch anzumerken, dass das Eliminieren von Beobachtungen immer auch mit dem Risiko verbunden ist, der Randomisierung entgegenzuwirken und dadurch Ergebnisse zu verzerren - insbesondere wenn die entfernten Beobachtungen ähnliche Charakteristiken tragen (wie z.B. viele Missing Values bzw. Missing Values an den selben Stellen) (Puma et al., 2009). Ebenfalls ist die Abwägung, ab welchem Punkt die Quote an Missing Values zu groß ist, sehr subjektiv.

Fehlende Werte können jedoch auch selbst eine Aussagekraft haben, weshalb es ebenfalls sinnvoll sein kann, diese beizubehalten. In dem Fall kann ein fehlender Wert wie eine zusätzliche Ausprägung der Variable betrachtet werden (Allison et al., 2010). Im Rahmen dieser Arbeit wird ein Mischverfahren angewandt - Beobachtungen in denen keine Werte für die abhängigen Variablen vorliegen, werden eliminiert. Missing Values an anderen Stellen werden aufgefüllt. Es ist anzumerken, dass Werte zu Variablen der Umfrage nur vorhanden sein können, wenn die entsprechende Person auch für die Umfrage ausgewählt wurde, weshalb darauf gefiltert werden muss.

Das Data Cleaning besteht aus mehreren aufeinander folgenden Schritten. Der Übersichtlichkeit halber, wurde für jeden einzelnen Teilschritt eine eigene Funktion geschrieben, welche am Ende mit *pandas.pipe()* mittels einer Pipeline auf den Datensatz angewandt werden können. Die Funktionen sind so geschrieben, dass sie universell auf jede der drei Umfragen bzw. die daraus entstandenen Datensätze angewandt werden können. Da die Variablennamen in den jeweiligen

Umfragen immer einen Suffix entsprechend der Kennung der Umfrage (0m, 6m und 12m) haben, muss der Funktion der entsprechende Suffix überreicht werden, damit Umformungen an spezifischen Spalten bzw. Variablen vorgenommen werden können. Im Folgenden werden die einzelnen Funktionen, also die Teilschritte des Data Cleanings erklärt:

```
def cache_dependant_variables(dataset:pd.DataFrame,
 name_suffix:str):
df = dataset.copy(deep=True)
df = df[df["returned" + name_suffix] == "Yes"]
cache = pd.DataFrame()
cache["health_gen" + name_suffix] = df["health_gen" +
 name_suffix].copy(deep=True)
cache["happiness" + name_suffix] = df["happiness" +
 name_suffix].copy(deep=True)
cache["med_qual" + name_suffix] = df["med_qual" +
 name_suffix].copy(deep=True)
return cache
```

Die Funktion *cache_dependant_variables* kopiert den zunächst unbereinigten Input-Datensatz und speichert die Columns der abhängigen Variablen in einem weiteren DataFrame zwischen, weil mit diesen anders verfahren wird, als mit den übrigen Variablen. Letztere werden später einem Dummy Encoding unterzogen, was mit den abhängigen Variablen nicht geschehen soll. Der Output der Funktion sind die zwischengespeicherten Variablen, welche im späteren Verlauf der Pipeline einer anderen Funktion übergeben werden.

```
def select_survey_participants(dataset:pd.DataFrame,
 name_suffix:str):
df = dataset.copy(deep=True)
df = df[df["returned" + name_suffix] == "Yes"]
return df
```

Im Schritt *select_survey_participants* wird der unbereinigte Datensatz übergeben und ausschließlich auf Zeilen begrenzt, welche Personen enthalten, welche die Umfrage beantwortet haben. Für Personen, die entweder nicht Teil des jeweiligen Samples waren, oder nicht auf die Umfrage geantwortet haben, liegen keine Werte für die Variablen der Umfrage vor, weshalb diese nicht für die Untersuchung der kausalen Effekte geeignet sind. Die Funktion gibt den neuen Datensatz, welcher nur aus Umfrageteilnehmern besteht, zurück.

```
def map_treatment_variables(df:pd.DataFrame,
 name_suffix:str):
treatment = {"Selected": 1, "Not selected": 0}
treatment2= {"Yes": 1, "No": 0}
df["treatment"] = df["treatment"].replace(treatment).
 astype(int)
df["treatment2"] = df["approved_app"].replace(
 treatment2).fillna(0).astype(int)
return df
```

Die Funktion *map_treatment_variables* wandelt die beiden Spalten mit den Treatment-Variablen vom String-Format in ein numerisches Format um. Die Spalte **treatment** steht dabei dafür, ob eine Person in der Lotterie ausgewählt wurde, während die ab hier „treatment2“ genannte Spalte angibt, wessen Bewerbung für MedicAid genehmigt wurde. Also gibt **treatment** die Intention to treat an, während **treatment2** die eigentliche Treatment-Variable im Kontext der Kausalen Inferenz widerspiegelt. Im ersten Schritt werden die Fälle, in denen ein Treatment erhalten wurde, mit einer 1 überschrieben, alle anderen Fälle mit einer 0. Bei der Variable **treatment2** gilt das auch für potenziell vorhandene Missing Values, da diese gleichbedeutend damit sind, dass die Bewerbung nicht akzeptiert wurde. Die Funktion gibt den Datensatz mit den aktualisierten Treatment-Variablen zurück.

```
def drop_problematic_columns(df:pd.DataFrame,
 name_suffix:str):
df.drop(["sample" + name_suffix, "dt_notify_lottery","
 dt_mail" + name_suffix], axis=1, inplace=True)
df.drop(["health_gen" + name_suffix, "happiness" +
 name_suffix, "med_qual" + name_suffix], axis=1,
 inplace=True)
return df
```

Durch *drop_problematic_columns* werden zum einen Spalten entfernt, die beim Survey-übergreifenden Preprocessing zu Problemen führen (primär wegen nicht uniformer Nomenklatur) und entweder bereits durch andere Variablen abgedeckt werden, oder augenscheinlich keinen Erklärungsbeitrag leisten. Im nächsten Schritt werden die abhängigen Variablen aus dem Datensatz entfernt, damit diese nicht das weitere Preprocessing unterlaufen - die Columns werden später mithilfe der bereits vorgestellten Funktion *cache_dependant_variables* wieder hinzugefügt.

```
def preprocess_object_columns(df:pd.DataFrame):
dataset_objects = df.select_dtypes(exclude=['number', '
 datetime'])
dataset_numbers = df.select_dtypes(include='number')
# split object columns in binary columns and non-binary
  columns
x = dataset_objects.apply(lambda x: x.nunique() ==2)
bool_cols = x[x==True].index
y = dataset_objects.apply(lambda x: x.nunique() !=2)
non_bool_cols = y[y==True].index
# get dummies for object columns
binary = pd.get_dummies(dataset_objects[bool_cols],
 dummy_na =True, drop_first=True).astype(int)
non_binary = pd.get_dummies(dataset_objects[
 non_bool_cols], dummy_na=True, drop_first=False).
 astype(int)
# merge the clean object and number dataframes
clean_data = pd.concat([dataset_numbers, binary,
 non_binary], axis=1)
return clean_data
```

Durch die Funktion *preprocess_object_columns* wird eine Trennung in numerische und kategoriale Variablen vorgenommen. Die kategorialen Variablen können gerade in Form von „Ja/Nein“-Fragen häufig in binärer Form vorliegen. *pandas* bietet mit der Funktion *select_dtypes()* die Möglichkeit, alle Spalten eines oder mehrerer ausgewählter Datentypen zu filtern. Da der Aufbau der Umfragen in Bezug auf die abgefragten Variablen identisch ist, sind die Schritte zur Umformung der Variablen ebenfalls (bis auf die Nomenklatur einzelner Variablen) identisch. Die numerischen Variablen können unverändert in das Modell eingehen - die kategorialen Variablen werden weiterverarbeitet. Zunächst wird in binäre und nicht-binäre Variablen getrennt. Theoretisch ist dieser Schritt nicht zwingend notwendig, bietet aber die Option mit binären Variablen anders zu verfahren. Im nächsten Schritt werden Dummy-Variablen erzeugt. Das bedeutet, dass die verschiedenen Ausprägungen der Variablen

auf mehrere Spalten aufgeteilt werden. In dem einfachen Fall, dass eine Spalte *X* die Ausprägungen „Ja“ und „Nein“ enthält, würden zwei neue Spalten angelegt werden: *X_Ja* und *X_Nein*. Für alle Zeilen, in denen in der ursprünglichen Spalte der Wert „Ja“ vorlag, wird in der neuen Spalte *X_Ja* eine 1 und in der Spalte *X_Nein* eine 0 eingetragen und vice versa. Dieses Verfahren ist skalierbar auf eine beliebige Anzahl an Antwortmöglichkeiten *k* - der limitierende Faktor ist dabei Rechenleistung und Modellperformance, welche unter zu viel Dummy-Encoding leiden kann (Fradkin & Madigan, 2003).

Um mit Dummy-Variablen alle Antwortmöglichkeiten abzubilden, werden *k-1* Spalten benötigt (Suits, 1957). Im Falle von binär verteilten Variablen, wie auch dem vorigen Beispiel, ist eine Spalte also ausreichend, um alle Antwortmöglichkeiten abzudecken. Wird nur die Spalte *X_Ja* betrachtet, bedeutet eine 0, dass in der ursprünglichen Spalte *X* ein „Nein“ stand. Da ein Großteil der Variablen jedoch auch noch *NaN* enthält und diese beibehalten werden sollen, gibt es faktisch drei Ausprägungen bei den „binären“ Variablen des Datensatzes. Mithilfe von *pandas*’ Funktion *get_dummies()* werden. Die Funktion *preprocess_object_colums* erstellt Dummy-Variablen für alle übergebenen Spalten. Für die Variable **binary** werden alle binären Variablen übergeben, für die Variable **non-binary** die restlichen. Mit *dummy_na=True* wird der Funktion *get_dummies()* übergeben, dass *NaN*-Werte ebenfalls codiert werden sollen. Standardmäßig wird für jede Ausprägung einer Variable eine Dummy-Spalte erzeugt - da theoretisch jedoch nur *k-1* Spalten notwendig sind, kann mit *drop_first=True* die erste Spalte gelöscht werden. In dieser Arbeit wurde sich dazu entscheiden, *drop_first=True* für Binärvariablen zu aktivieren, um den bereinigten Datensatz bezüglich der Dimensionalität geringer zu halten. Für nicht-binäre Variablen wurde *drop_first=False* gesetzt.

Zuletzt müssen die nun allesamt numerischen Datensatz-Bestandteile wieder zusammengefügt werden. Die *pandas*-Funktion *concat()* ermöglicht es, mehrere Datensätze zusammenzufügen. Über die Achse wird angegeben, ob die Datensätze untereinander oder nebeneinander verbunden werden sollen - *axis=1* bedeutet, dass die Datensätze nebeneinander gesetzt werden. Häufig wird zum Zusammensetzen von DataFrames eine gemeinsame Variable als Schlüssel benötigt. In diesem Fall sind die Datensätze jedoch alle in der identischen Reihenfolge, in der sie sich auch befunden haben, bevor sie getrennt wurden. Deshalb kann in diesem Fall ohne einen spezifizierten Schlüssel verbunden werden. *pandas* verwendet, falls nicht weiter spezifiziert, automatisch den Index. Der Output der Funktion ist ein Datensatz, welcher im klassischen Sinne bereinigt ist und lediglich Spalten im numerischen Datentyp enthält.

```
def fill_missing_values(clean_data:pd.DataFrame):
for column in clean_data.columns:
    if clean_data[column].isna().values.any():
        clean_data[column + "wasNaN"] = clean_data[
 column].isna() * 1
clean_data.fillna(0, inplace=True)
return clean_data
```

Da der Datensatz nach dem vorigen Schritt noch Missing Values enthält, muss abgewägt werden, wie damit verfahren werden soll. Wie bereits zuvor in der Arbeit erklärt, werden die Missing Values in diesem Fall aufgefüllt. Die fehlenden Werte lagen fast ausschließlich in Spalten vor, in denen eine numerische Angabe von den Teilnehmern gefragt war - andere Spalten derselben Reihen waren sowohl vor, als auch nach den fehlenden Werten trotzdem gefüllt. Es gibt verschiedene Gründe, warum Teilnehmer Fragen in einer Umfrage unbeantwortet lassen. Beispielsweise, weil sie selbst nicht genau wissen, welche die akkurate Antwort ist oder weil die Frage für sie zu privat ist. Ebenfalls können Fragen unbeantwortet bleiben, wenn Teilnehmer nach einigen Fragen die Umfrage abbrechen. Die Variablen bzw. Fragen, welche im OHIE eine vergleichsweise hohe Quote an Missing Values aufweisen, sind vermehrt Fragen nach der Häufigkeit. Diese Fragen sind für Versuchsteilnehmer im Kontext von MedicAid thematisch zu erwarten gewesen, weshalb hier nicht davon ausgegangen wird, dass die Fragen aufgrund von Bedenken hinsichtlich der Privatsphäre nicht beantwortet wurden. Gerade bei Fragen nach der Häufigkeit könnte ein Nicht-Beantworten der Antwort „0“ gleichgesetzt worden sein. Deshalb werden durch die Funktion *fill_missing_values* sämtliche verbleibende Missing Values im Datensatz mit 0 aufgefüllt. Da jedoch durchaus die Möglichkeit besteht, dass die fehlenden Werte mehr Bedeutung hatten, werden zuvor sämtliche Spalten, welche Missing Values enthalten um eine weitere Spalte bzw. Variable erweitert. Diese zusätzliche Spalte kennzeichnet, wo zuvor in dem ursprünglichen Column Missing Values vorhanden waren, damit der potenzielle Erklärungsgehalt der fehlenden Werte erhalten bleibt.

Theoretisch wird durch das Auffüllen der Missing Values ein Bias eingeführt. Da jedoch die fehlenden Werte im Datensatz weiterhin gekennzeichnet bleiben, ist die Idee dieser Arbeit, dass der durch das Auffüllen entstandene Bias geringer ist als der, der entstünde, wenn sämtliche Reihen mit Missing Values eliminiert und der Datensatz um ein Vielfaches verkleinert werden würde.

```
def add_and_transform_dependant_columns(clean_data:pd.
 DataFrame, cache:pd.DataFrame, name_suffix:str):
health = {"excellent": 4, "very good": 3, "good": 2, "
 fair": 1, "poor": 0}
happiness = {"very": 2, "pretty happy": 1, "not too
 happy": 0}
quality = {"excellent": 4, "very good": 3, "good": 2, "
 fair": 1, "poor": 0, "no care": 5}
clean_data["med_qual_" + name_suffix + "nocare"] = (
 cache["med_qual" + name_suffix] == "no care") * 1
clean_data["health_gen" + name_suffix] = cache["
 health_gen" + name_suffix].copy(deep=True).map(health)
clean_data["happiness" + name_suffix] = cache["
 happiness" + name_suffix].copy(deep=True).map(
 happiness)
clean_data["med_qual" + name_suffix] = cache["med_qual"
  + name_suffix].copy(deep=True).map(quality)
med_qual_mean = clean_data[clean_data["med_qual" +
 name_suffix] != 5]["med_qual" + name_suffix].dropna().
 astype(int).mean()
```

```
clean_data["med_qual" + name_suffix] = clean_data["
 med_qual" + name_suffix].replace(5, med_qual_mean)
return clean_data
```

Im Anschluss an das Auffüllen der Missing Values werden durch die Funktion
add_and_transform_dependant_columns die zuvor entfernten abhängigen Variablen wieder mit dem Datensatz verbunden. Dafür werden der Funktion der bereinigte Datensatz und die Cache-Funktion übergeben. Die Columns werden ebenfalls mit der *map()*-Funktion von *pandas* in die numerische Form überführt. Eine Besonderheit liegt beim Column **med_qual** vor. Bei diesem wird die Qualität der in Anspruch genommen medizinischen Leistungen beurteilt. Jedoch besteht auch die Möglichkeit, dass ein Teilnehmer keine Leistungen in Anspruch genommen hat. Dieser Fall wurde im ursprünglichen Datensatz mit „no care“ gekennzeichnet. In der Funktion wird der Wert „no care“ zunächst mit einem Platzhalter überschrieben und anschließend durch den Mittelwert aller Werte, die nicht „no care“ waren, ersetzt. Hintergrund ist, dass dadurch ein großer Anteil an Zeilen im Datensatz erhalten werden kann. Das Auffüllen mit dem Mittelwert verzerrt zwar die Ergebnisse, aber weniger, als wenn ein Wert von größer als 4 beziehungsweise kleiner als 0 eingesetzt werden würde. Zudem ist zu beachten, dass der Wert „no care“ lediglich überschrieben wird, wenn **med_qual** als unabhängige Variable verwendet wird. In dem Fall, dass **med_qual** als unabhängige Variable betrachtet wird, werden die Zeilen mit dem Mittelwert entfernt. Der Output der Funktion ist der bereinigte Datensatz, welcher um die nun numerischen abhängigen Variablen erweitert wurde.

```
def drop_remaining_missing_values(clean_data:pd.
 DataFrame):
clean_data.dropna(axis=0, inplace=True)
return clean_data
```

Da die in der vorherigen Formel wieder hinzugefügten Spalten das Preprocessing übersprungen haben, enthalten diese noch Missing Values, welche durch die Funktion
drop_remaining_missing_values entfernt werden.

```
def transform_remaining_columns_to_int(clean_data:pd.
 DataFrame, name_suffix:str):
clean_data["health_gen" + name_suffix]  = clean_data["
 health_gen" + name_suffix].astype(int)
clean_data["happiness" + name_suffix] = clean_data["
 happiness" + name_suffix].astype(int)
clean_data["med_qual" + name_suffix]  = clean_data["
 med_qual" + name_suffix].astype(int)
return clean_data
```

Zuletzt müssen die unabhängigen Variablen noch mit der Funktion
transform_remaining_columns_to_int in das Integer-Format überführt werden, da diese nach dem Mapping in *pandas* immer noch als Object-Column interpretiert werden.

```
clean_data_12m_filled =(
df_12.pipe(select_survey_participants, "_12m")
.pipe(map_treatment_variables, "_12m")
.pipe(drop_problematic_columns, "_12m")
.pipe(preprocess_object_columns)
.pipe(fill_missing_values)
.pipe(add_and_transform_dependant_columns,
 cache_dependant_variables(df_12, "_12m"), "_12m")
.pipe(drop_remaining_missing_values)
.pipe(transform_remaining_columns_to_int, "_12m")
)
```

Nun kann ein vollständig bereinigter Datensatz erzeugt werden, indem der ursprüngliche Datensatz mit der Funktion *pipe()* schrittweise den zuvor vorgestellten Funktionen übergeben wird. In jedem Schritt ist erkennbar, welcher Teil des Preprocessings angewandt wird und welche zusätzlichen Inputs, wie beispielsweise der Names-Suffix, übergeben werden. Der Prozess wurde in Abbildung 7 visualisiert dargestellt.

Da in dem weiteren Verlauf der Arbeit zwei unterschiedliche Datensätze verwendet werden, muss zu dem bereinigten Datensatz in voller Größe noch ein Datensatz erzeugt werden, bei dem auf die Variable **numhh** bedingt wird, um ein vollständig randomisiertes Experiment sicherzustellen.

```
clean_data_12m_cond = clean_data_12m_filled[
 clean_data_12m_filled["numhh_list_signed self up"] ==
 1]
```

Die Ausprägung „signed self up“ ist die am häufigsten vorkommende, weshalb auf diese Ausprägung bedingt wird, um den größtmöglichen Datensatz beizubehalten. Um Double Machine Learning mithilfe der *DoubleML*-Bibliothek auf den Datensatz anzuwenden, muss dieser in dem *DoubleMLData*-Format vorliegen. Dafür wird die Funktion *DoubleMLData()* verwendet. Als erstes wird der *pandas*-DataFrame übergeben, anschließend werden die Spalten bzw. Variablen spezifiziert. Es muss mindestens ein Treatment übergeben werden und es besteht die Option Instrumentalvariablen zu übergeben. Exemplarisch sieht dies wie folgt aus:

```
obj_dml_data = DoubleMLData(dataframe, y_col="outcome",
  d_cols="treatment", z_cols="IV_variables", x_cols="
 covariates")
```

Die resultierende Python-Variable **obj_dml_data** liegt in einer Form vor, mit der Double Machine Learning betrieben werden kann.

4.4 RandomForest

Wie im Unterkapitel zu den kausalen Modellen bereits erwähnt, wird in dieser Arbeit RandomForest als Machine Learning Algorithmus verwendet. RandomForest ist ein sogenannter Ensemble-Learner - also eine Kombination mehrerer Algorithmen, die eine höhere Leistung erzielt als die jeweils einzelnen Methoden (Breiman, 2001). In der Regel wird versucht Modelle zu kombinieren, die sich gegenseitig ergänzen bzw. die Schwächen der anderen Modelle ausgleichen. Im Falle von RandomForest werden zwei Verfahren miteinander kombiniert: Decision Trees und Bagging.

Decision Trees sind aufgrund ihrer intuitiven Interpretierbarkeit weit verbreitet. Decision Trees splitten den Datensatz anhand der Variablen so auf, dass die Schätzungen - gemessen an einer zuvor ausgewählten Metrik - verbessert werden. Dieser Vorgang nennt sich Recursive Partitioning. Um

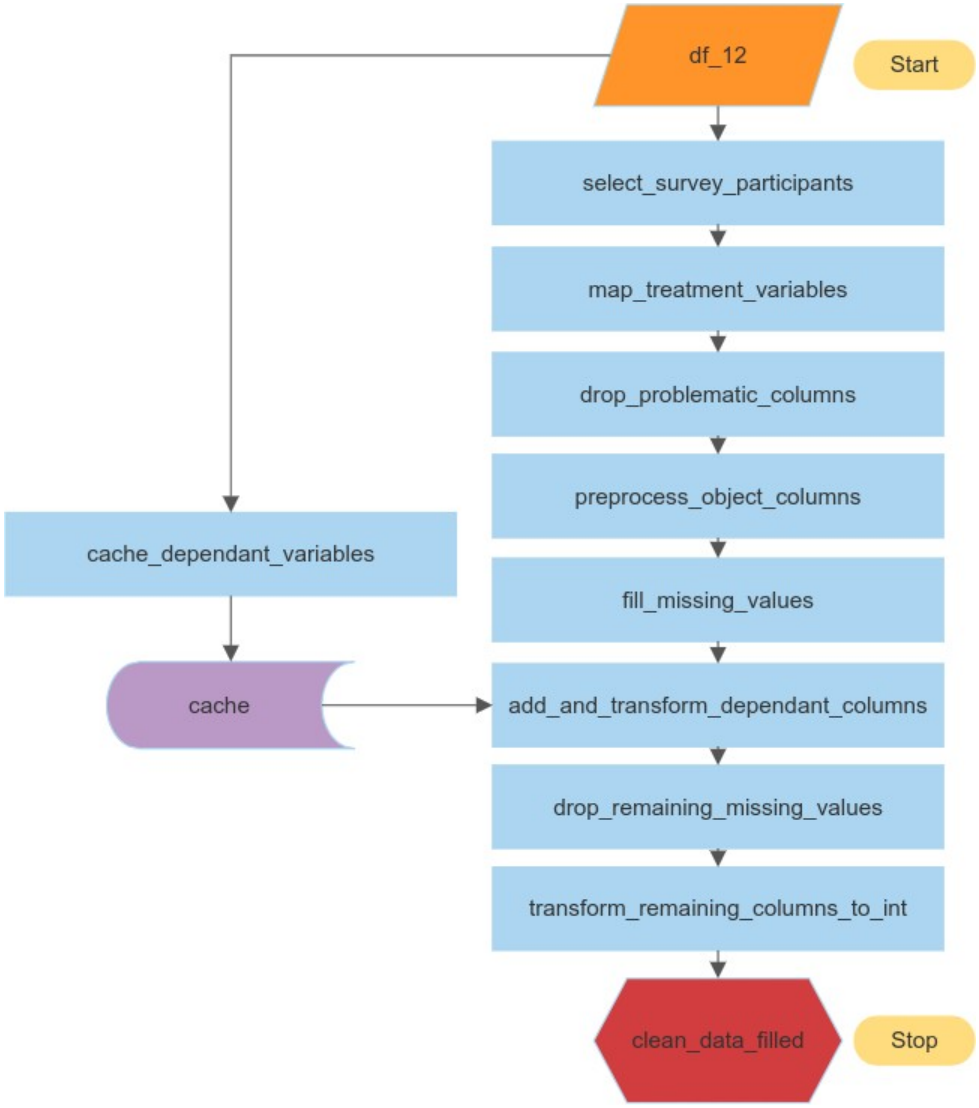

Abbildung 7: Preprocessing Flowchart (Quelle: E. D.)

einen Decision Tree zu errichten, wird Recursive Partitioning über mehrere Iterationen hinweg wiederholt - in jeder Iteration wird der Split gewählt, welcher den Schätzungsfehler am meisten reduziert (von Winterfeld & Edwards, 1986). Dieser Prozess kann theoretisch wiederholt werden, bis jeder Datenpunkt einen eigenen Knoten im Decision Tree bildet. Der Vorteil von Decision Trees ist ersichtlich: Ihre einfache Erstellung und ihre leichte Interpretierbarkeit. Der Nachteil des Modells ist jedoch, dass mit tieferen Bäumen die Wahrscheinlichkeit zunimmt, die Daten zu overfitten (Bramer, 2007). Im Extremfall, wenn jeder einzigartige Datenpunkt einen eigenen Knoten im Decision Tree hat, werden die Daten vollständig auswendig gelernt, und die Varianz des Modells ist hoch - dann wäre zu erwarten, dass die Performance auf Testdaten nicht zufriedenstellend ist.

Bagging - ausgeschrieben: Bootstrap Aggregating - ist ein Verfahren, das die Performance von Machine-Learning-Algorithmen verbessert und gleichzeitig die Varianz des Modells reduziert, indem es Overfitting verhindert. Die zentrale Idee von Bagging besteht darin, mehrere Sample aus dem Trainingsdatensatz zu ziehen, auf denen anschließend Modelle trainiert werden. Diese Modelle werden dann zu einem neuen, robusteren Modell aggregiert (Breiman, 1996). Die Besonderheit beim Ziehen der Sample ist, dass dies „mit Zurücklegen“ geschieht. Ein besonderes Merkmal des Verfahrens ist, dass die Stichproben mit Zurücklegen gezogen werden. Das bedeutet, dass jede Beobachtung potenziell in mehreren Stichproben vorkommen kann. Es ist möglich, dass einige Beobachtungen mehrfach gezogen werden, während andere gar nicht gezogen werden.

Der RandomForest-Algorithmus ist eine Kombination dieser beiden Methoden - aus dem Trainingsdatensatz werden n Sample gezogen, auf die jeweils ein Decision Tree trainiert wird. Anschließend werden diese Decision Trees zu einem Modell aggregiert. Das finale Modell ist robuster als einzelne Decision Trees (Syrgkanis & Zampetakis, 2020). Ran-

domForest hat zudem viele Hyperparameter, die optimiert werden können. Im Folgenden werden die Hyperparameter vorgestellt, welche im weiteren Verlauf der Arbeit mittels Hyperparameter-Tuning angepasst werden.

Der Hyperparameter ***n_estimators*** gibt an, wie viele Entscheidungsbäume in den RandomForest einbezogen werden und gehört damit zu den wichtigsten Parametern des Verfahrens. Grundsätzlich gilt, dass mit zunehmender Anzahl an Entscheidungsbäumen die Präzision der Ergebnisse steigt, bzw. diese robuster werden - der Grenzertrag nimmt mit steigender Zahl an Decision Trees ab. Mit ***max_features*** wird festgelegt, wie viele Variablen maximal pro Split an einem Knoten betrachtet werden, während ***max_depth*** die maximale Tiefe beziehungsweise Knotenzahl eines Entscheidungsbaumes angibt. Die maximale Anzahl an Features, die an einem Knoten betrachtet wird, erhöht die Variabilität der einzelnen Entscheidungsbäume zueinander. Wenn ***max_features*** kleiner ist als die Anzahl der Variablen im Datensatz n, wird nur ein zufälliges Subset an Variablen für jeden Split berücksichtigt. Ein geringeres ***max_features*** als n kann daher dazu beitragen, dass die Schätzungen des Random Forest robuster werden. Mit ***max_depth*** hingegen wird Overfitting reduziert, indem die Entscheidungsbäume auf eine maximale Tiefe begrenzt werden, über die sie nicht hinauswachsen können. Ist der Parameter gering genug gesetzt, wird verhindert, dass die Decision Trees jede einzelne Beobachtung auswendig lernen. Der Hyperparameter ***min_samples_leaf*** gibt an, wie viele Beobachtungen mindestens in einem Endpunkt eines Entscheidungsbaumes vorhanden sein müssen. Je nachdem, wie hoch der Wert gesetzt wird, verhindert auch dies Overfitting, da beispielsweise nicht für jede Beobachtung ein einzelner Endpunkt erzeugt werden kann. Obwohl beide Hyperparameter dazu dienen Overfitting zu reduzieren, ist es sinnvoll beide zu optimieren, da die Wirkungsweise sich unterscheidet. Zum Beispiel kann es vorkommen, dass ein Modell auf einem großen Datensatz bereits lange overfittet, aber ***min_samples_leaf*** noch nicht auslöst, da immer noch genügend Beobachtungen in jedem Endknoten enthalten sind. Der Parameter ***max_depth*** hätte möglicherweise den Entscheidungsbaum bereits abgebrochen und vice versa (Cournapeau, n. d.).

4.5 Hyperparameter-Tuning

Im Bereich des Machine Learnings gibt es bei der Arbeit mit Algorithmen häufig verschiedene Variablen, die innerhalb eines Algorithmus angepasst werden können. Diese Variablen, auch Hyperparameter genannt, beeinflussen den Lernprozess des Algorithmus und somit auch dessen Qualität (Schratz et al., 2019). Das sogenannte Hyperparameter-Tuning (manchmal auch Hyperparameter-Optimization) beschreibt den Prozess diese Parameter so auszuwählen, dass die Performance des Modells auf den individuellen Datensatz optimiert wird. Es gibt keinen allgemeingültigen „besten“ Wert für Hyperparameter. Es gibt jedoch einen theoretisch fundierten Default-Wert, der standardmäßig als Ausgangspunkt und somit als Baseline verwendet wird. Dieser Wert liefert in vielen Fällen auch ohne Hyperparameter-Tuning bereits zufriedenstellende Ergebnisse und rechtfertigt somit seine Verwendung als Default (Wong et al., 2019). Hyperparameter-Tuning ist ein Prozess, bei dem speziell für den Anwendungsfall optimale Werte für Hyperparameter bestimmt werden. Im Rahmen dieser Arbeit soll Hyperparameter-Tuning auf den zuvor erklärten RandomForest-Algorithmus angewendet werden, um so präzise Ergebnisse wie möglich zu erhalten. Im folgenden Unterkapitel werden zwei Verfahren erläutert, welche für das Hyperparameter-Tuning relevant sind.

4.5.1 Crossvalidation und Gridsearch

Nachdem zuvor bereits einige der optimierbaren Hyperparameter vorgestellt wurden, sollen im folgenden Abschnitt die Algorithmen beschrieben werden, die zur Optimierung der Hyperparameter verwendet werden. Ein gängiger Ansatz ist die Crossvalidation, bei der der Datensatz in ein Trainings- und Testset aufgeteilt wird. Eine Möglichkeit dies umzusetzen, ist die *k-fold Crossvalidation*. Dabei wird der Datensatz in k Subsets unterteilt - k kann dabei vom Anwender festgelegt werden. In jeder Iteration werden *k-1* Subsets als Trainingsdatensatz und das verbleibende Subset als Testset verwendet (Bro et al., 2008). Dabei werden in jeder Iteration verschiedene Hyperparameter-Einstellungen verwendet. Nach k Durchläufen werden die Modelle hinsichtlich ihrer Performance miteinander verglichen, um die besten Hyperparameter-Einstellungen zu ermitteln (Probst et al., 2019).

Wenn eine große Anzahl von Hyperparametern optimiert werden soll, die wiederum eine Vielzahl von möglichen Ausprägungen haben können, kann die Anzahl der zu testenden Modelle einen immensen Rechenaufwand darstellen. Hat ein Algorithmus beispielsweise lediglich einen Hyperparameter, welcher binär verteilt ist, gäbe es zwei Modelle welche gegeneinander getestet werden können. Bei fünf Parametern, welche jeweils im Bereich von 0 und 10 verteilt sein könnten, ist die Anzahl verschiedener Modelle bereits 161.051 Stück. Jedes dieser Modelle zu trainieren und zu validieren ist ein großer Rechenaufwand. Darüber hinaus ist es oft gar nicht sinnvoll, jedes denkbare Modell zu testen, da kleine Unterschiede in den Hyperparametern möglicherweise keinen signifikanten Einfluss auf die Modellleistung haben. In der Praxis ist es oftmals nicht relevant, ob 300 oder 301 Decision Trees in einem RandomForest trainiert werden (Probst et al., 2019). In diesem Fall kann das *Gridsearch*-Verfahren verwendet werden, um bestimmte Werte für die Hyperparameter auszuwählen, die gegeneinander getestet werden sollen. Werden für mehrere Parameter Werte angegeben, so werden die Kombinationen miteinander getestet und es ergibt sich ein Netz - daher auch der Name (Hsu et al., 2003). Die Werte der einzelnen Hyperparameter für *Gridsearch* können entweder manuell ausgewählt werden oder es kann ein Intervall gesetzt werden, aus welchem dann zufällig Werte gezogen werden - in diesem Fall wird von *Randomsearch* gesprochen

(Bergstra & Bengio, 2012).

4.5.2 Implementation

Im Folgenden wird die Umsetzung der beschriebenen Verfahren zum Hyperparameter-Tuning von RandomForest in Python dargestellt. Das *DoubleML*-Package hat eine eingebaute Funktion *tune()*, welche zum Tunen der Hyperparameter verwendet werden kann. Es besteht die Möglichkeit zwischen Gridsearch und Randomsearch zu wählen. Der Funktion wird ein Parameter-Grid mit allen zu testenden Ausprägungen der Hyperparameter übergeben. Nachfolgend wird die Anwendung der *tune()*-Funktion exemplarisch dargestellt.

```
ml_g = RandomForestRegressor()
ml_m = RandomForestClassifier()

param_grid = {
'n_estimators': np.array([150, 300]),
'max_depth': np.array([25, None]),
'max_features': np.array(["sqrt", None])
}

par_grids = {'ml_g': param_grid,
             'ml_m': param_grid}

obj_dml_irm.tune(par_grids, search_mode='grid_search')
fit = obj_dml_irm.fit()
```

Im obigen Code-Block werden zunächst die verwendeten Machine Learning Algorithmen definiert. Anschließend wird das Parameter-Grid definiert, welches die verschiedenen Ausprägungen der Hyperparameter enthält, die gegeneinander getestet werden sollen. Anschließend wird das Parameter Grid jedem der Machine Learner übergeben. Nachdem die Parameter fürs Tuning gesetzt wurden, muss das Grid lediglich innerhalb der *tune()*-Funktion abgerufen werden. Innerhalb der Funktion wird der Search-Algorithmus festgelegt - in diesem Fall *Gridsearch*. Abschließend wird die *fit()*-Funktion aufgerufen und der *Gridsearch*-Algorithmus wird automatisch während des Fitting-Prozesses des Modells ausgeführt.

Für diese Arbeit wurden die Parameter ***n_estimators***, ***max_depth*** und ***max_features*** als die wichtigsten Parameter ausgewählt. Da die benötigte Rechenzeit für das Hyperparameter-Tuning stark mit der Anzahl der Parameter wächst, ist schnell ein Limit erreicht, insbesondere bei großen Datensätzen. Um die Rechenanforderungen zu senken, wurde bei den IRM ein kleineres Parameter-Grid verwendet, als bei den IIVM. In der Tabelle 1 ist eine Auflistung der Werte zu sehen, welche den Modellen für die jeweiligen Hyperparameter übergeben wurden.

Alle Ausprägungen des Parameter-Grids für Interactive Regression Models finden sich auch im Grid für die Interactive IV Models wieder - letztere haben pro Hyperparameter jedoch noch jeweils eine weitere Ausprägung.

Hyperparameter	**Werte für IRM**	**Werte für IIVM**
n_estimators	300	300, 350
max_depth	25, None	10, 25, None
max_features	sqrt, None	log2, sqrt, None

Tabelle 1: Vergleich der Parameter-Grids fürs Hyperparameter-Tuning (Quelle: E. D.)

4.6 Untersuchung kausaler Effekte

Die Untersuchung der kausalen Effekte im nachfolgenden Kapitel ist von wesentlicher Bedeutung für die Beantwortung der Leitfrage dieser Arbeit, welche Effekte Public Healthcare auf die abgedeckte Bevölkerung eines Landes hat und wie stark diese sind. Nachdem in den vorangegangenen Kapiteln die theoretischen Grundlagen gelegt wurden, erfolgt in dem fünften Kapitel die Untersuchung der Auswirkungen des Treatments auf verschiedene Outcome-Variablen. Davor wird in diesem Abschnitt der dafür verwendete Code dargestellt.

Das Vorgehen für jede untersuchte Variable ist identisch: Zunächst wird der kausale Effekt der Intention to treat untersucht, d.h. wie stark der Einfluss der alleinigen Absicht ist, ein MedicAid-Programm durchzuführen. Dies dient als Baseline für die weiteren Schätzungen - sollten der Treatment-Effekt sich nicht von der Intention to treat unterscheiden, scheint es keinen signifikanten Treatment-Effekt zu geben. Dafür wird ein Interactive Regression Model verwendet. Zum Vergleich wird dasselbe Modell verwendet, um den eigentlichen Treatment-Effekt, d.h. die Aufnahme in das MedicAid-Programm, zu untersuchen. Nach Abschluss dieser beiden Untersuchungen wird der Treatment-Effekt erneut unter Verwendung des Interactive-IV-Regression-Modells geschätzt, wobei die Auswahl in der Lotterie als Instrumentalvariable verwendet wird. Dieser Ablauf wird pro Variable zweimal durchgeführt - auf jedem der Datensätze einmal. Es werden also für jede Variable, die Ergebnisse für die folgenden Fälle ermittelt:

- Datensatz: Missing Values aufgefüllt - (Vollständig)
  - Intention to treat - IRM
  - Treatment Effekt - IRM
  - Treatment Effekt - IIVM
- Datensatz: Missing Values aufgefüllt - (Bedingt auf *numhh*)
  - Intention to treat - IRM
  - Treatment Effekt - IRM
  - Treatment Effekt - IIVM

Nachfolgend ist exemplarisch der verwendete Code aufgeführt. Für jeden geschätzten kausalen Effekt wurde diese Struktur verwendet - Unterschiede gibt es lediglich bei der betrachteten Outcome-Variable, dem verwendeten Modell und der Treatment-Variable.

```
obj_dml_data = DoubleMLData(data=DataFrame, y_col=
 outcome_variable, d_cols=treatment_variable, z_cols=
 instrumental_variable)
obj_dml_iivm = DoubleMLIIVM(obj_dml_data=obj_dml_data,
 ml_g=ml_g, ml_m=ml_m, ml_r=ml_r, apply_cross_fitting=
 True, n_folds=2)

param_grid = {
    'n_estimators': np.array([300, 350]),
    'max_depth': np.array([10, 25, None]),
    'max_features': np.array(["log2", "sqrt", None])
}
par_grids = {'ml_g': param_grid,
             'ml_m': param_grid,
              'ml_r': param_grid}

obj_dml_iivm.tune(par_grids)
fit = obj_dml_iivm.fit()
print(fit.summary)
```

In der ersten Zeile wird das *DoubleMLData*-Objekt mithilfe der Funktion *DoubleMLData()* erzeugt. Die Funktion hat mehrere Parameter, die gesetzt werden müssen. Dem Parameter *data* wird der verwendete Datensatz übergeben, dem Parameter *y_col* die Spalte für die unabhängige Variable und dem Parameter *d_cols* die Treatment-Variable. Falls, wie im exemplarischen Beispiel, ein Interactive IV Model verwendet wird, muss zusätzlich eine Instrumentalvariable an den Parameter *d_cols* übergeben werden. In der zweiten Zeile wird das Modell ausgewählt, das verwendet werden soll. Im exemplarischen Beispiel wurde das IIVM verwendet. Die zu übergebenden Parameter sind das *DoubleML*-Data Objekt und die zu verwendenden Machine Learner. Zusätzlich kann angegeben werden, ob Crossfitting angewandt werden soll und in wie viele Samples der Datensatz dafür aufgeteilt werden soll. Wann immer eine zufällige Aufteilung der Daten in Samples geschieht, kann ein sogenannter Seed festgelegt werden. Unter Verwendung des gleichen Seeds wird immer dieselbe Aufteilung der Daten erreicht werden. So kann ermöglicht werden, Ergebnisse später zu reproduzieren.

5 Ergebnisse

In diesem Kapitel werden die Ergebnisse der Schätzungen der kausalen Effekte präsentiert und ausgewertet, welche im Zuge der Untersuchung der abhängigen Variablen zur Beantwortung der Leitfragen gewonnen wurden. Jeder Abschnitt dieses Kapitels widmet sich einer der untersuchten abhängigen Variablen. Dabei werden in jedem Abschnitt jeweils zwei Abbildungen dargestellt - eine für jeden Datensatz. Die Ergebnisse werden im Fließtext auf vier und in den Abbildungen auf sechs Nachkommastellen gerundet. Im ersten Abschnitt wird die wahrgenommene Gesundheit der Teilnehmer betrachtet, im zweiten Abschnitt die Anzahl der Arztbesuche innerhalb der letzten sechs Monate und im dritten Abschnitt die Zufriedenheit der Teilnehmer. Die beiden darauf folgenden Abschnitte untersuchen, ob der Bedarf an medizinischen Leistungen und verschreibungspflichtigen Medikamenten gedeckt wurde. Anschließend wird noch die Qualität der erhaltenen medizinischen Leistungen untersucht. Der letzte Abschnitt fasst die Ergebnisse in Bezug auf die wissenschaftliche Literatur zum Thema zusammen und vergleicht sie mit den Ergebnissen anderer Studien zum Oregon Health Insurance Experiment.

5.1 Wahrgenommene Gesundheit

Die Variable **health_gen_12m** gibt den selbstgeschätzten Gesundheitszustand der Teilnehmer wieder - die Skala gab dabei die Optionen „exzellent“, „sehr gut“, „gut“, „okay“ und „schlecht“ her, aus denen die Personen den jeweils zutreffendsten Wert auswählen konnten. Da für die Arbeit mit den Modellen die String-Werte in numerische Werte überführt werden mussten, wurden den einzelnen Ausprägungen Zahlenwerte von 0-4 zugewiesen, wobei 4 für den Gesundheitszustand „exzellent“ und eine 0 für den Zustand „schlecht“ steht. Die Betrachtung dieser Variable sollte dazu beitragen, die Leitfrage, inwieweit MedicAid einen Einfluss auf die Gesundheit der Probanden hat, zu beantworten. Zudem wurde die Variable bereits von anderen Studien untersucht, weshalb die Möglichkeit besteht, Ergebnisse zu vergleichen. Im folgenden werden die Ergebnisse der verschiedenen Durchläufe aufgezeigt:

Bei der Schätzung der kausalen Effekte auf den vollständigen Datensatz (Abbildung 8) ist zu erkennen, dass die Intention to treat einen geringen Koeffizienten hat. Mit einer Höhe von 0,0092 ist der Koeffizient so niedrig, dass der Effekt vernachlässigbar gering zu sein scheint. Wird der Treatment-Effekt geschätzt, so ergeben sich mit einem Interactive Regression Model ein Wert von 0,0282 und mit dem Interactive IV Model ein Wert von 0,0476. Es fällt auf, dass die Wirkungsrichtung beider Effekte positiv ist und beide Effekte auf dem 1%-Niveau statistisch signifikant sind.

Ein Vergleich der vorigen Ergebnisse mit denen aus dem randomisierten Datensatz (Abbildung 9) zeigt, dass die Effekte in allen drei untersuchten Szenarien sehr ähnlich geschätzt wurden. Der Effekt der Intention to treat liegt bei 0,0113, während die Koeffizienten für den Treatment-Effekt vom IRM auf 0,0269 und vom IIVM auf 0,0488 geschätzt wurden. Auch hier sind alle Ergebnisse signifikant, wie am t-Test gesehen werden kann. Da der vollständige Datensatz streng genommen kein randomisiertes Experiment darstellt, sollte die Schätzung des Interactive IV Models am robustesten sein. Bei dem randomisierten Datensatz sollte die Verwendung einer Instrumentalvariable nicht notwendig sein - im Falle von Störfaktoren kann es jedoch trotzdem sinnvoll sein Instrumentalvariablen zu verwenden. An dieser Stelle lässt sich nicht mit Sicherheit sagen, welches Modell den kausalen Effekt besser schätzt weshalb dieser als Intervall zwischen den Schätzungen des IRM und IIVM angegeben wird. Damit liegt der kausale Koeefizient zwischen 0,0269 und 0,0488. Zwar ist die Wirkungsrichtung des Koeffizienten positiv, doch selbst das obere Ende des Intervalls ist mit einem Koeffizienten in Höhe von 0,0488 so gering, dass vermutlich nicht von einem relevanten kausalen Effekt gesprochen werden kann. Es scheint zwar einen Wirkungszusammenhang zwischen dem Treatment und **health_gen_12m** zu geben, doch in Anbetracht der Tatsache, dass die Skala der

health_gen_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.009184	0.003328	2.759999	0.00578	0.002662	0.015707
Treatment Effekt - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.028206	0.003353	8.412796	4.00E-17	0.021635	0.034777
Treatment Effekt - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.047588	0.01606	2.963045	0.003046	0.01611	0.079065

Abbildung 8: Kausale Effekte „Wahrgenommene Gesundheit“ - A (Quelle: E. D.)

health_gen_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.011298	0.003925	2.878413	0.003997	0.003605	0.018992
Treatment Effekt - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.026924	0.004365	6.167512	6.94E-10	0.018368	0.03548
Treatment Effekt - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.048759	0.017578	2.773913	0.005539	0.014307	0.083211

Abbildung 9: Kausale Effekte „Wahrgenommene Gesundheit“ - B (Quelle: E. D.)

Gesundheit von 0-4 reicht, ist eine Verbesserung des Gesundheitszustandes um 0,0488 bei Erhalt des Treatments nicht von Bedeutung.

Im Zuge der Beantwortung der Leitfrage konnte im Rahmen der Untersuchung zwar ein positiver Kausalzusammenhang zwischen MedicAid und wahrgenommener Gesundheit festgestellt werden, jedoch ist dieser vernachlässigbar gering. Es sind viele Szenarien denkbar, wie das Ergebnis der Analyse erklärt werden kann. Beispielsweise bedeutet die Möglichkeit dazu eine Behandlung finanziert zu bekommen nicht automatisch, dass ein Patient diese in Anspruch nimmt. Es wäre denkbar, dass Patienten mit ihren Beschwerden einen Arzt aufsuchen, im Anschluss aber die empfohlenen Therapie-Möglichkeiten nicht wahrnehmen. Der Datensatz des OHIE bietet keine Möglichkeit zu erkennen, ob Patienten tatsächlich Behandlungen bekommen und kontinuierlich bis zum Abschluss wahrgenommen haben. Es sollte also für möglich gehalten werden, dass MedicAid in diesem Fall nicht das Treatment, sondern die Option zum Erhalt eines Treatments ist. Wenn dem so wäre, könnte der gemessene kausale Effekt von MedicAid auf die Gesundheit von Menschen wieder als eine Art „Intention to treat“-Effekt betrachtet werden. Was zusätzlich berücksichtigt werden sollte ist zudem, dass die Beurteilung des Gesundheitszustandes eine Selbsteinschätzung der Probanden war. Es ist also nicht sichergestellt, dass alle Teilnehmer des OHIE das gleiche Verständnis der potenziellen Antwortoptionen zum eigenen Gesundheitszustand hatten, da diese subjektive Kriterien waren. Es ist denkbar, dass diese Abweichungen sich im Mittel ausgleichen, jedoch wäre es für zukünftige Arbeiten sinnvoller ein objektives Kriterium für die Abbildung des Gesundheitszustandes der Probanden zu bilden. Die gerechtfertigtste Interpretation ist an dieser Stelle, dass der Sachverhalt nicht mit Sicherheit korrekt interpretiert werden kann und zusätzliche Daten gesammelt und untersucht werden sollten, um mehr Aufschluss zu geben.

Diese Ergebnisse stimmen zu einem gewissen Grad mit der bisherigen Studienlage zum OHIE überein. Finkelstein et al. schreiben in ihrem Paper „The Oregon Health Insurance Experiment: Evidence from the First Year“, dass versicherte

doc_num_mod_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.011647	0.015533	0.74978	0.453387	-0.018798	0.042091
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.151387	0.015568	9.724253	2.38E-22	0.120875	0.1819
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.078637	0.075736	1.038302	0.299129	-0.069803	0.227078

Abbildung 10: Kausale Effekte „Anzahl Arztbesuche“ - A (Quelle: E. D.)

doc_num_mod_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.021015	0.01946	-1.079902	0.280186	-0.059156	0.017126
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.192995	0.019247	10.027286	1.16E-23	0.155272	0.230719
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.020602	0.086781	0.237405	0.812343	-0.149486	0.190691

Abbildung 11: Kausale Effekte „Anzahl Arztbesuche“ - B (Quelle: E. D.)

Individuen eine um 25% höhere Wahrscheinlichkeit haben, ihren Gesundheitszustand mit „gut“, „sehr gut“ oder „exzellent“ zu bewerten, als die Kontrollgruppe (Finkelstein et al., 2012). Dies deutet auf einen positiven Zusammenhang zwischen Versicherung und guter Gesundheit hin. Baicker et al. hingegen fanden in ihrer Publikation „The Oregon Experiment — Effects of MedicAid on Clinical Outcomes“ keinen signifikanten Effekt von MedicAid auf die physische Gesundheit (Baicker et al., 2013). Die Ergebnisse dieser Arbeit stimmen hinsichtlich der Wirkungsrichtung mit dem von Finkelstein et al. ermittelten Effekt überein, schätzen diesen jedoch deutlich geringer ein. Der durch Double Machine Learning ermittelte kausale Effekt in Höhe von 0,0269-0,0488 wurde als vernachlässigbar gering eingeschätzt, was mehr mit den Ergebnissen von Baicker et al. übereinstimmt. Es kann angenommen werden, dass die Ergebnisse unter Anwendung von Double Machine Learning eine höhere Validität aufweisen, da diese den tatsächlichen kausalen Effekt ermitteln, anstatt Assoziationen zu belegen.

5.2 Anzahl der Arztbesuche

Die Variable **doc_num_mod_12m** steht für die Anzahl an Arztbesuchen, welche die Teilnehmer innerhalb der letzten sechs Monate getätigt haben. Diese Variable lag bereits in numerischer Form vor, also war kein Data-Cleaning neben dem Umgang mit Missing Values nötig. Im Folgenden sind die Ergebnisse der Untersuchung:

Bei der Schätzung des kausalen Effekts von MedicAid auf die Anzahl an Arztbesuchen ist erkennbar, dass die Ergebnisse der Intention to treat und des IIVM nicht statistisch signifikant sind, da der t-Test das 1%-Signifikanzniveau überschreitet (Abbildung 10). Lediglich der Koeffizient des IRM in Höhe von 0,1514 ist signifikant. Die Ergebnisse des vollständigen Datensatzes weisen also auf einen positiven Zusammenhang zwischen Treatment und Outcome hin. Unter Betrachtung der Ergebnisse auf dem randomisierten Datensatzes (Abbildung 11) fällt auf, dass wieder lediglich die Schätzung des Interactive Regression Models mit einem Koeffizienten von 0,193 signifikant ist. Da die anderen Koeffizienten das 1%-Signifikanzniveau nicht unterschreiten, werden diese nicht betrachtet. Sowohl bei dem vollständigen Daten-

satz, als auch bei dem auf **numhh** bedingten Datensatz, ist die Wirkungsrichtung der signifikanten Koeffizienten positiv.

An den Ergebnissen ist erkennbar, dass ein positiver kausaler Effekt vorliegt von dem Treatment auf die Anzahl der Arztbesuche vorliegt. Zwar sind die Ergebnisse zu der Intention to treat und die des IIVM nicht signifikant, jedoch haben die Schätzungen des Interactive Regression Models auf beiden Datensätzen signifikante Schätzungen hervorgebracht. Da das IRM nicht dafür geeignet ist, mit Nicht-randomisierten Daten zu arbeiten, wird die Schätzung des Koeffizienten auf dem randomisierten Datensatz als zuverlässiger betrachtet (Bach et al., 2021). Es wird an dieser Stelle also interpretiert, dass der kausale Koeffizient von MedicAid auf die Anzahl an Arztbesuchen bei 0,193 liegt.

Demnach lässt sich interpretieren, dass MedicAid-Coverage dazu führte, dass die Teilnehmer der Studie öfter Ärzte aufgesucht haben. Eine Abdeckung durch MedicAid erhöht durchschnittlich die Anzahl an Arztbesuchen um etwa 0,193 mal. Während dieser Effekt für ein Individuum isoliert betrachtet gering zu sein scheint, ist zu berücksichtigen, dass der Effekt über viele Personen hinweg ins Gewicht fallen kann. Geht jede von 100 Personen im Schnitt zusätzliche 0,193 mal zum Arzt, sind das etwa 19 zusätzliche Arztbesuche. Eine mögliche Erklärung dafür ist, dass ein Arztbesuch ohne Versicherung in der Regel mit Kosten verbunden ist. Gerade in den USA fallen Kosten für medizinische Leistungen sehr hoch aus, weshalb potenziell der Trade-off zwischen Kosten und Dringlichkeit einer Behandlung beziehungsweise eines Arztbesuches abgewogen wird (Vladeck, 2003). Durch die Abdeckung eines MedicAid-Programms, welches im Falle eines Besuches oder einer medizinischen Prozedur die Kosten trägt, muss dieser Trade-off nicht mehr gemacht werden - die wirtschaftliche Hemmschwelle einen Arzt zu besuchen könnte also abnehmen.

Diese Ergebnisse decken sich mit denen von Baicker et al., welche zeigen, dass versicherte Individuen häufiger Ärzte besuchen, als die Kontrollgruppe. Sie fanden einen Unterschied im Mittelwert zwischen Versuchs- und Kontrollgruppe von 2,7 (Baicker et al., 2013). Der Effekt konnte auch in dieser Arbeit belegt werden, wenn auch deutlich geringer (0,193). Auch an dieser Stelle ist davon auszugehen, dass die Ergebnisse unter Verwendung von Double Machine Learning robuster und somit zuverlässiger sind.

5.3 Zufriedenheit

In der Variable **happiness_12m** wurde die selbst wahrgenommene Zufriedenheit der Teilnehmer erfasst. Ähnlich wie bei der Variable zum Gesundheitszustand gab es hier ebenfalls eine Item-Skala, bei der die Teilnehmer ihre Zufriedenheit mit den Attributen „sehr glücklich“, „glücklich“ und „nicht glücklich“ bewerten konnten. Auch hier wurden die Strings im Preprocessing in numerische Werte, in diesem Fall von 0 bis 2 überführt, wobei 2 für „sehr glücklich“ steht. Die Grundzufriedenheit der Bevölkerung ist eine wichtige Variable, die sich auf viele weitere Faktoren auswirkt. Generell gilt, dass eine höhere Zufriedenheit der Menschen sowohl auf politischer und gesellschaftlicher, als auch auf individueller Ebene angestrebt wird (Malthus, 1888). Im Folgenden sind die Ergebnisse der Untersuchung aufgeführt:

Die Koeffizienten zum kausalen Effekt von MedicAid auf die Zufriedenheit der Teilnehmer sind im ersten Datensatz allesamt signifikant (Abbildung 12). Der Effekt der Intention to treat und des IRM sind mit 0,0212 und 0,0322 ähnlich hoch. Der vom Interactive IV Model geschätzte Koeffizient liegt bei 0,1227 und ist damit deutlich höher.

Die Ergebnisse ähneln den Ergebnissen vom randomisierten Datensatz (Abbildung 13). Auch hier sind alle Koeffizienten signifikant. Die Intention to treat hat einen Koeffizienten von 0,0292 und IRM und IIVM haben einen Koeffizienten von 0,0292 bzw. 0,1189. Alle kausalen Effekte haben dabei eine positive Wirkungsrichtung.

Auf dem vollständigen Datensatz haben die Schätzungen des IIVM die höchste theoretische Validität. Damit lässt sich die Obergrenze des kausalen Effektes auf 0,1227 schätzen. Die Schätzung des IRM ist nur auf dem randomisierten Datensatz valide, wo sie sich jedoch nicht signifikant von der Schätzung der Intention to treat unterscheidet. Wie zuvor erwähnt, kann nur von einem kausalen Effekt gesprochen werden, wenn der geschätzte Effekt und die Intention to treat verschieden voneinander sind. Auch hier wird wieder ein Intervall für den kausalen Koeffizienten geschätzt. Dieser liegt zwischen 0,0292 und 0,1227, wobei die untere Intervallgrenze nahezu identisch mit der Intention to treat ist. Anhand des Intervalls lässt sich kein relevanter Effekt bzw. ein positiver Effekt von MedicAid-Coverage auf die Zufriedenheit der Empfänger interpretieren. Im Verhältnis zur Größe der Skala ist dieser jedoch gering - der Koeffizient macht selbst bei Verwendung der oberen Intervallgrenze nur etwa ein Achtel des Sprunges zur nächsten Stufe aus. Theoretisch wäre denkbar gewesen, dass die Sorge um medizinische Not-Situationen und potenziell hohe Kosten die Zufriedenheit der Nicht-Versicherten mindern beziehungsweise die Zufriedenheit der Versicherten durch die Auflösung dieser Sorge steigern - diese Vermutung lässt sich empirisch anhand der Daten des Oregon Health Insurance Experiments zumindest teilweise bestätigen beziehungsweise nicht vollständig widerlegen. Ob weniger Sorge um medizinische Kosten tatsächlich der Grund ist, oder noch ein anderer Aspekt dahintersteht, müsste in Zukunft näher untersucht werden.

Der Einfluss von MedicAid auf die Zufriedenheit wurde auch von Baicker et al. untersucht. Diese fanden keinen signifikanten Effekt von MedicAid auf die Zufriedenheit der Versuchsgruppe (Baicker et al., 2013). Da im Rahmen dieser Arbeit ein Intervall für den kausalen Effekt geschätzt wurde, ist der Vergleich nur begrenzt möglich. Unter Verwendung der unteren Intervallgrenze von 0,0292 lässt sich der Effekt als vernachlässigbar gering einschätzen und stimmt somit mit den Ergebnissen von Baicker et al. überein. Die Ver-

happiness_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.021169	0.003375	6.273152	3.54E-10	0.014555	0.027783
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.032173	0.003365	9.560503	1.17E-21	0.025578	0.038769
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.122665	0.016633	7.374804	1.65E-13	0.090065	0.155265

Abbildung 12: Kausale Effekte „Zufriedenheit“ - A (Quelle: E. D.)

happiness_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.029178	0.003969	7.350671	1.97E-13	0.021398	0.036958
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.029196	0.003971	7.353267	1.93E-13	0.021414	0.036978
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.118856	0.01778	6.684848	2.31E-11	0.084008	0.153704

Abbildung 13: Kausale Effekte „Zufriedenheit“ - B (Quelle: E. D.)

wendung der oberen Intervallgrenze von 0,1227 hingegen deutet auf einen kausalen Zusammenhang zwischen Treatment und Outcome hin, wenn auch einen sehr geringen. An dieser Stelle wird darauf verwiesen, dass zukünftig weitere Untersuchungen bzw. Studien den Zusammenhang zwischen Gesundheitsversicherung und Zufriedenheit untersuchen sollten, bevor eindeutige Schlussfolgerungen gezogen werden können.

5.4 Medizinischer Bedarf gedeckt

Ob ein Teilnehmer sämtliche, innerhalb der letzten sechs Monate benötigten, medizinischen Leistungen erhalten hat, wurde in der Variable **needmet_med_cor_12m** erfasst. Die Befragten konnten angeben, dass sie entweder keine Leistungen benötigten, oder diese alle erhalten haben - in diesem Fall steht in der bereinigten Spalte eine 1. Hat der Teilnehmer nicht alle benötigten medizinischen Leistungen erhalten, ist dies durch eine 0 gekennzeichnet. Die Frage danach, ob sämtlicher Bedarf an medizinischen Leistungen einer Person gedeckt werden kann, ist eine wichtige. Sind bestimmte medizinische Leistungen für manche Menschen aufgrund ihrer finanziellen Situation nicht erschwinglich, dann stellt dies ein ethisches Problem dar (Summers & Morrison, 2009). Ob die Kosten medizinischer Leistungen tatsächlich einen Einfluss darauf haben, ob diese bei Bedarf in Anspruch genommen werden, sollte untersucht werden. Nachfolgend sind die Ergebnisse der Analyse abgebildet:

Die geschätzten kausalen Koeffizienten, die MedicAid auf den Erhalt benötigter medizinischer Leistungen hat, sind im vollständigen Datensatz allesamt nicht signifikant (Abbildung 14). Die Ergebnisse im bedingten Datensatz sind ebenfalls nicht signifikant und werden daher nicht weiter betrachtet (Abbildung 15).

An dieser Stelle lässt sich lediglich die Annahme äußern, dass Behandlungen, welche tatsächlich notwendig sind, von den Betroffenen Personen trotz hohen Kosten dennoch in Anspruch genommen werden. Theoretisch lässt sich hier auch ein Fall von Survivorship Bias vermuten - Personen die wirklich notwendige Behandlungen trotz hoher Kosten in Anspruch nehmen, haben vermutlich eine höhere Chance im Datensatz aufzutauchen, als Menschen, die dringende Behandlungen aus finanziellen Gründen nicht in Anspruch nehmen (können), da verstorbene Personen aus dem Datensatz entfernt wurden. Dies sind jedoch nur Vermutungen, welche

needmet_med_cor_12m_Yes						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.000045	0.000044	1.014984	0.310114	-0.000042	0.000131
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.00015	0.000094	-1.596246	0.110434	-0.000334	0.000034
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.000207	0.000295	0.701115	0.483231	-0.000371	0.000785

Abbildung 14: Kausale Effekte „Bedarf Med. Leistungen“ - A (Quelle: E. D.)

needmet_med_cor_12m_Yes						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.000072	0.000055	1.305576	0.191697	-0.000036	0.000181
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.000212	0.000125	1.702199	0.088718	-0.000032	0.000457
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.000188	0.000217	0.867655	0.385583	-0.000237	0.000614

Abbildung 15: Kausale Effekte „Bedarf Med. Leistungen“ - B (Quelle: E. D.)

an dieser Stelle nicht mit empirischen Belegen aus den Ergebnissen dieser Arbeit unterstützt werden können.

Ob MedicAid einen Einfluss darauf hat, inwieweit der Bedarf an medizinischen Leistungen gedeckt wird, wurde bereits von Baicker et al. im Paper „The Effect of MedicAid on Management of Depression“ untersucht. Sie fanden, dass durch MedicAid die Wahrscheinlichkeit, dass der Bedarf an medizinischen Leistungen nicht gedeckt wird, um 10,7% reduziert wurde (Baicker et al., 2018). Diese Ergebnisse können anhand der Untersuchung dieser Arbeit weder empirisch belegt, noch widerlegt werden. Es wäre in Zukunft interessant, ähnliche Daten und damit auch diese Fragestellung, genauer zu untersuchen.

5.5 Bedarf an Medikamenten gedeckt

Die Variable **needmet_rx_cor_12m** misst, ob der Bedarf eines Studienteilnehmers an verschreibungspflichtigen Medikamenten innerhalb der letzten sechs Monate gedeckt wurde. Auch hier steht eine 1 dafür, dass keine Medikamente benötigt oder alle benötigten Medikamente erhalten wurden, während eine 0 angibt, dass nicht alle Bedürfnisse gedeckt wurden. Das Konzept dieser Fragestellung ähnelt der des vorherigen Unterabschnitts, jedoch werden medizinische Behandlungen und Medikamente im Rahmen dieser Arbeit als zwei unterschiedliche Ausprägungen betrachtet. Beispielsweise können die Kosten für die Behandlung chronischer Erkrankungen, die eine langfristige medikamentöse Therapie erfordern, im Vergleich zu den einmaligen Kosten einer Operation erheblich höher und fortlaufend sein (Hussey et al., 2014). Diese potenzielle Unterscheidung war Grund dafür die Variable **needmet_rx_cor_12m** separat zu untersuchen. Die Ergebnisse sind im Folgenden zu sehen:

Bei der Untersuchung des vollständigen Datensatzes ist lediglich der Koeffizient des Interactive Regression Models signifikant (Abbildung 16). Dieser beträgt -0,0003 und ist zwar negativ, aber kann approximativ als 0 betrachtet und vernachlässigt werden. Bei der Schätzung auf dem vollständigen Datensatz ist das IRM weniger geeignet, als das IIVM, allerdings kann die generelle Größenordnung des Koeffizienten angenommen werden.

Bei den Ergebnissen auf dem randomisierten Datensatz ist erneut nur die Schätzung des Interactive Regression Mo-

needmet_rx_cor_12m_Yes						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	0.000084	0.000052	1.629699	0.103165	-0.000017	0.000185
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.000321	0.000114	-2.810998	0.004939	-0.000544	-0.000097
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.00028	0.000307	0.913443	0.36101	-0.000321	0.000881

Abbildung 16: Kausale Effekte „Bedarf Medikamente“ - A (Quelle: E. D.)

needmet_rx_cor_12m_Yes						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.000049	0.000074	-0.656675	0.51139	-0.000193	0.000096
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.001031	0.000215	-4.788631	0.000002	-0.001453	-0.000609
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.000301	0.000334	-0.901043	0.367566	-0.000955	0.000354

Abbildung 17: Kausale Effekte „Bedarf Medikamente“ - B (Quelle: E. D.)

dels signifikant (Abbildung 17). Der geschätzte Koeffizient beträgt hier -0,001 und ist ebenfalls negativ und approximativ 0. An dieser Stelle lässt sich interpretieren, dass MedicAid-Coverage im Rahmen der Daten des Oregon Health Insurance Experiments keinen messbaren kausalen Effekt auf die Abdeckung des Bedarfs an Medikamenten hatte. Es lässt sich vermuten, dass notwendige Medikamente unabhängig von den finanziellen Aspekten in Anspruch genommen werden. An dieser Stelle wird jedoch ausdrücklich darauf hingewiesen, dass die empirische Evidenz in diesem Abschnitt begrenzt ist und weitere Untersuchungen notwendig sind, um diese Hypothese zu bestätigen. Im Zuge dessen wäre es interessant, weitere Variablen wie z.B. finanzielle Aspekte in die Betrachtung einzuziehen.

Bisher wurde noch keine wissenschaftliche Untersuchung darüber durchgeführt, inwiefern MedicAid den Bedarf der Empfänger an verschreibungspflichtigen Medikamenten deckt. Deshalb ist ein direkter Vergleich mit bisheriger empirischer Evidenz nicht möglich. Baicker et al. publizierten jedoch in ihrem Paper „The Effect of MedicAid on Management of Depression“, dass MedicAid zu einem erhöhten Einsatz von Medikamenten bei der Kontrollgruppe führte (Baicker et al., 2018). Theoretisch wäre denkbar, dass ein vermehrter Einsatz von Medikamenten auch bedeutet, dass der Bedarf an Medikamenten besser gedeckt wird. Andererseits ist ebenso vorstellbar, dass der Mehr-Einsatz von Medikamenten kaum Überschneidung mit dem tatsächlichen Bedarf hat und überwiegend „optionale“ Medikamente verwendet werden. Da die Ergebnisse dieser Arbeit die Kausalbeziehung nicht näher beleuchten, bedarf es in Zukunft weiterer Untersuchungen, um den Zusammenhang zu klären.

5.6 Qualität medizinischer Leistungen

Die letzte untersuchte Variable ist **med_qual_12m**. Diese Variable erfasst, inwieweit die Teilnehmer die Qualität der in den letzten sechs Monaten in Anspruch genommenen medizinischen Leistungen beurteilen. Hier wurde ebenfalls wieder eine Item-Skala verwendet, welche im Preprocessing in numerische Form umgewandelt wurde. Die Skala bestand aus den Attributen „exzellent“, „sehr gut“, „gut“, „in Ordnung“, „schlecht“ und „keine Leistungen beansprucht“. Der Skala wurden die Werte 0 bis 4 vergeben,

med_qual_12m						
Missing Values gefüllt - ganzer Datensatz						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.007299	0.005358	-1.362208	0.173132	-0.0178	0.003203
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.003905	0.005393	-0.724185	0.468952	-0.014475	0.006664
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.024027	0.02328	-1.032073	0.302038	-0.069655	0.021602

Abbildung 18: Kausale Effekte „Qualität medizinischer Leistungen“ - A (Quelle: E. D.)

med_qual_12m						
Missing Values gefüllt - auf numhh bedingt						
Intention to treat						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment	-0.014055	0.006442	-2.18187	0.029119	-0.026681	-0.001429
Treatment Effect - IRM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	0.001862	0.00634	0.293614	0.769053	-0.010565	0.014289
Treatment Effect - IIVM						
	coef	**std err**	**t**	**P>|t|**	**2.50%**	**95.50%**
treatment2	-0.052689	0.025323	-2.080666	0.037465	-0.102322	-0.003057

Abbildung 19: Kausale Effekte „Qualität medizinischer Leistungen“ - B (Quelle: E. D.)

wobei „exzellent“ als Wert 4 gesetzt worden ist. Wenn die Variable **med_qual_12m** als Kovariable verwendet wurde, wurde „keine Leistungen beansprucht“ mit dem Mittelwert aufgefüllt - damit der potenziell vorhandene Erklärungsgehalt jedoch nicht verloren geht, wurde in einer neu erzeugten binären Spalte festgehalten, welche Teilnehmer keine Leistungen in Anspruch genommen haben. Für die Untersuchung innerhalb dieses Abschnitts wurden zusätzlich zu allen fehlenden Werten in **med_qual_12m** zudem auch die Teilnehmer aus dem Datensatz entfernt, die keine medizinischen Leistungen innerhalb der letzten sechs Monate beansprucht haben. Damit soll sichergestellt werden, dass kein Bias eingeführt wird.

Die Ergebnisse auf dem vollständigen Datensatz sind allesamt nicht auf dem 1%-Niveau signifikant und werden daher nicht weiter betrachtet (Abbildung 18). Die Schätzungen auf dem randomisierten Datensatz sind, ebenfalls nicht auf dem 1%-Niveau signifikant (Abbildung 19). Die Intention to treat und das Interactive IV Model sind jedoch auf dem 5%-Niveau statistisch signifikant. Der Koeffizient der Intention to treat beträgt -0,0141 und der des IIVM beträgt -0,0527. Damit ist der geschätzte Effekt von der Intention to treat verschieden - es scheint also einen negativen kausalen Zusammenhang zwischen MedicAid und der Qualität der erhaltenen medizinischen Leistungen zu geben. In Anbetracht der Größe der Skala lässt sich jedoch wieder feststellen, dass der Effekt vernachlässigbar gering ist. Ebenfalls ist hier anzumerken, dass wie bereits in Kapitel 5.1 mit einer subjektiven Skala gearbeitet wurde. Insbesondere auf die Bewertung von medizinischen Leistungen können viele subjektive Faktoren einwirken, die nicht zwangsläufig mit der Qualität der Leistung zusammenhängen. Ebenfalls sollte angemerkt werden, dass der durchschnittlichen Bevölkerung die fachlichen Qualifikationen fehlen könnten, um aussagekräftige Bewertungen medizinischer Leistungen treffen zu können.

Im Rahmen dieser Arbeit lässt sich also interpretieren, dass MedicAid sich nicht merklich auf die Qualität der erhaltenen medizinischen Leistungen auswirkt, die empirische Evidenz jedoch gering ist. Aus ethischen Gründen sollte jeder Patient unabhängig von seiner Versicherung bestmöglich behandelt werden (Summers & Morrison, 2009). Es wäre jedoch von großem Interesse zu untersuchen, ob diese Idealvorstellung in der Realität umgesetzt wird. Die Varia-

ble **med_qual_12m** wurde bislang in keiner anderen Publikation zum Oregon Health Insurance Experiment untersucht, was keinen Vergleich empirischer Evidenz ermöglicht.

5.7 Einordnung OHIE

Die Menge der statistisch signifikanten Ergebnisse, welche aus dieser Arbeit hervorgehen, ist limitiert. Eine direkte Überschneidung der Ergebnisse mit denen anderer Studien zum OHIE besteht nur bei drei untersuchten Variablen (siehe Kapitel 5.1-5.3). Ein Vergleich dieser Ergebnisse mit anderen Studien zeigt, dass es sowohl Übereinstimmungen als auch Unterschiede gibt. Die Übereinstimmung gibt es primär in der Wirkungsrichtung der geschätzten Effekte. Unterschiede gibt es jedoch in ihrer Größe - es zeigte sich (in der limitierten Samplegröße $n = 3$) dass Double Machine Learning tendenziell geringere Effektgrößen geschätzt hat, als klassische statistische Methoden. Eine Einordnung in die bisherigen Ergebnisse zum Oregon Health Insurance Experiment lässt sich wie folgt formulieren: Die Untersuchungen kausaler Effekte durch MedicAid auf verschiedene Variablen haben gezeigt, dass andere Studien zum OHIE Kausalzusammenhänge in ihrer Wirkungsrichtung akkurat beleuchten, jedoch in der Größe der Effekte nicht akkurat bestimmen konnten. Da im Rahmen des OHIE viele weitere Faktoren und Variablen untersucht wurden, wäre es interessant diese Studien ebenfalls mit Double Machine Learning zu untersuchen und potenziell zu reevaluieren. Gerade wenn die Ergebnisse des Oregon Health Insurance Experiments Basis für zukünftige Entscheidungen zum Thema Gesundheitsversicherung bilden könnten, sollte aus wissenschaftlichem und generellem Interesse sichergestellt werden, dass die zugrundeliegenden Ergebnisse so akkurat wie möglich sind. Zukünftige Untersuchungen mit Double Machine Learning auf dem Datensatz des Oregon Health Insurance Experiments könnten beispielsweise zur mentalen Gesundheit der Teilnehmer durchgeführt werden - in diesem Bereich liegen bereits ausführliche Ergebnisse durch Baicker et al. vor, welche signifikante Effekte von MedicAid auf die Diagnose und Behandlung von Depressionen fanden (Baicker et al., 2018).

6 Kritik

Gerade bei der Auswertung von Daten, sollte jeder Schritt kritisch hinterfragt werden. Kleinste Fehler in der Datenerhebung, bei der Modellwahl oder bei anderen Prozessen, können einen Bias in das Ergebnis einführen und dieses damit verzerren. Laufen Prozesse beispielsweise, den äußeren Umständen geschuldet, nicht optimal ab, so ist es wichtig, darauf hinzuweisen. Dadurch soll verhindert werden, dass zukünftige Arbeiten Bezug auf die Ergebnisse nehmen und unbewusst den Bias weiterreichen. Zudem kann durch transparentes Arbeiten ein Bewusstsein für die bestimmte Arten von Fehlerquellen geschärft werden. In diesem Kapitel soll der gesamte Prozess, von Erhebung der Daten, bis hin zur Gewinnung der Ergebnisse kritisch betrachtet und auf potenzielle Fehlerquellen untersucht werden. Im folgenden Unterkapitel wird dabei zunächst der Datensatz des Oregon Health Insurance Experiments auf mögliche Fehlerquellen untersucht. Im zweiten Abschnitt wird die statistische Signifikanz der Ergebnisse betrachtet. Im dritten und letzten Abschnitt wird auf die Limitationen dieser Arbeit hingewiesen.

6.1 Datensatz

Während das OHIE ein großes Experiment von hoher Relevanz ist, lassen sich jedoch einige Kritikpunkte am Datensatz äußern. Im Rahmen dieser Arbeit werden die folgenden fünf Kritikpunkte am OHIE als die wichtigsten erachtet:

- Die Datensätze enthalten nicht genug Beobachtungen
- Es liegt potenziell ein Sampling Bias vor
- Es werden nicht alle Variablen von Interesse betrachtet
- Variablen unterliegen teilweise Subjektivität
- Die Daten spiegeln nur eine Momentaufnahme wieder

Ein wesentlicher Kritikpunkt ist die Samplegröße der jeweiligen Umfragen. Die Umfrage nach 12 Monaten, welche die Grundlage für die Analysen in dieser Arbeit war, enthielt nach dem Preprocessing noch 22.904 Datenpunkte im vollständigen und 16.064 Datenpunkte im randomisierten Datensatz. Organisatorische und finanzielle Gründe waren vermutlich limitierende Faktoren bei der Entscheidung der Samplegrößen der Umfragen, jedoch wäre es aus Datenperspektive besser gewesen, die Grundgesamtheit der Teilnehmer zu befragen.

Zusätzlich muss davon ausgegangen werden, dass es potenziell mehrere Quellen für einen Sampling Bias gibt. Im Falle des Oregon Health Insurance Experiments war die Auswahl der untersuchten Gesamtheit nicht zufällig - es wurden ausschließlich Menschen mit Wohnsitz in Oregon in das Sample aufgenommen. Während diese Tatsache in der Realität vermutlich keinen signifikanten Bias auf die Ergebnisse überträgt, sollte dennoch beachtet werden, dass theoretisch keine vollständige Übertragbarkeit der Ergebnisse gewährleistet wäre, falls sich Oregon signifikant von anderen Staaten in den USA unterscheidet. Wäre beispielsweise die medizinische Versorgung in Oregon signifikant schlechter, als in dem Rest Amerikas, würde das voraussichtlich kausale Effekte der öffentlichen Krankenversicherung auf die Gesundheit der Bevölkerung in ihrer Stärke beeinflussen oder möglicherweise sogar weitere Variablen durch Confounding betreffen. Gleiches gilt für die Übertragbarkeit auf andere Länder. Ebenfalls zu beachten ist, dass eine große Menge an Teilnehmern die Umfragen nicht beantwortet haben. Unterscheidet sich die Gruppe der Nicht-Antwortenden von den Teilnehmern, welche die Umfragen beantwortet haben, liegt ebenfalls ein Sampling Bias vor. Es wurden bereits monetäre Anreize gesetzt und Nachbearbeitungen durchgeführt, um die Antwortquote zu erhöhen, jedoch hätten diese potenziell höher bzw. intensiver sein müssen. In dieser Arbeit wird jedoch angenommen, dass der Effekt vernachlässigbar gering ist.

Ebenfalls kritisieren lässt sich, dass Variablen von potenziellem Interesse nicht im Datensatz erfasst wurden. Wie zuvor erwähnt besteht beispielsweise keine Möglichkeit in den Daten zu erkennen, ob Patienten tatsächlich Behandlungen bekommen und kontinuierlich bis zum Abschluss wahrgenommen haben. Gerade im Bereich der Kausalen Inferenz ist es von hoher Wichtigkeit, dass für den Kausalzusammenhang relevante Variablen nicht ausgelassen werden, da dies einen Bias einführen kann (Hernan & Robins, 2023). Zukünftig wäre es interessant mehr Variablen zu finanziellen Aspekten in den Datensatz einzubeziehen, da die Übernahme von Kosten ein wesentlicher Unterschied zwischen öffentlicher und privater Krankenversicherung ist (White, 2015).

Neben ausgelassenen Variablen lassen sich auch die erfassten Variablen kritisieren. Viele der Variablen wurden über subjektive Selbsteinschätzungen erhoben. Neben möglichen Faktoren, welche die Selbsteinschätzungen verzerrt haben könnten, ist es zudem denkbar, dass die Ausprägungen der Variable von den Teilnehmern unterschiedlich interpretiert wurden: was ein Proband als „exzellenter Gesundheitszustand“ einschätzt, ist immer relativ zur eigenen Wahrnehmung, nicht zu einem objektiven Standard. Für zukünftige Studien wäre es hier potenziell sinnvoll, zusätzlich zu subjektiven Variablen auch objektive Kenngrößen zu messen - im Falle der Gesundheit ließen sich beispielsweise Puls, Blutdruck, Sauerstoffsättigung, Gewicht, etc. ohne großen Mehraufwand ebenfalls bestimmen.

Zuletzt kann kritisiert werden, dass die Umfragen immer nur eine Momentaufnahme abbilden können. Zwar kann durch die Kombination der Umfragen ein längerer Zeitraum abgedeckt werden, jedoch lässt sich argumentieren, dass gerade bei langwierigen Themen wie Medizin und Gesundheit ein weitreichenderer Zeitraum hätte betrachtet werden müssen. Zusätzlich ist zu berücksichtigen, dass aufgrund der geringen Anzahl an Umfragen ein höheres Risiko besteht, dass die Umfrageergebnisse (insbesondere die subjektiven) durch äußere Umstände beeinflusst werden. Dadurch könnte die interne Validität der Studie eingeschränkt sein. Eine Erhöhung der Anzahl der Beobachtungen würde die Robustheit der Ergebnisse erhöhen und das Risiko von Verzerrungen durch äußere Einflüsse reduzieren (Hamaker et al., 2015).

Es lässt sich an dieser Stelle zusammenfassend sagen, dass das Oregon Health Insurance Experiment einen Datensatz hervorgebracht hat, aus dem sich erste Inferenzen bilden lassen - für eine vollständige Beleuchtung der Thematik sollte jedoch ein ähnliches Experiment erhoben werden, bei denen die potenziellen Fehler des OHIE ausgebessert werden.

6.2 Signifikanz der Ergebnisse

Die in dieser Arbeit gewonnenen Ergebnisse sind an vielen Stellen nicht statistisch auf dem 1%-Niveau (bzw. 5%-Niveau) signifikant gewesen. Während Double Machine Learning eine wertvolle Methode ist, um Kausale Inferenz anzuwenden, ist es nicht garantiert, dass es in allen Fällen statistisch signifikante Ergebnisse liefert. Es gibt zahlreiche Faktoren, die die statistische Aussagekraft der Ergebnisse beeinflusst haben könnten - unter anderem der Stichprobenumfang, das Preprocessing der Daten oder die Komplexität des zugrundeliegenden Kausalzusammenhangs (Croxton & Cowden, 1939; Hernan & Robins, 2023). Eine Wiederholung der Analyse mit anderer Aufteilung des Datensatzes (d.h. Random-Seed bei der Crossvalidation) führt cet. par. potenziell einen Survivorship Bias in die Ergebnisse ein, da lediglich gewünschte Ergebnisse betrachtet werden, weshalb davon in dieser Arbeit abgesehen wurde (Ioannidis, 2005). Generell ist davon abzuraten, statistisch nicht-signifikante Analysen zu wiederholen. Stattdessen sollten die angewandte Methodik und der zugrundeliegende Sachverhalt tiefergehend auf mögliche Fehlerquellen untersucht werden. Eine nicht-exklusive Liste der potenziellen Änderungen an den im Rahmen dieser Arbeit durchgeführten Untersuchungen beinhalten:

- anders mit Missing Values verfahren
- andere Machine Learning Algorithmen verwenden
- zusätzliches Hyperparamter-Tuning
- andere Variablen untersuchen
- Folds (n_folds) bei der Crossvalidation anders stratifizieren

Viele dieser Änderungen sind mit dem in dieser Arbeit bereitgestellten Code mit geringem Aufwand durchführbar. Für die Zukunft wäre es interessant zu sehen, ob die Signifikanz der Ergebnisse von der verwendeten Methodik beeinflusst wurde - im besten Fall lassen sich so langfristig „Best Practices“ im Umgang mit Double Machine Learning ermitteln, welche zukünftige Anwendungen in ihrer Genauigkeit verbessern könnten.

6.3 Limitationen

Eine Limitation bei der Erstellung dieser Arbeit war die benötigte Rechenleistung. Die Anwendung von Machine Learning und der Umgang mit großen Datenmengen ist meist rechnerisch anspruchsvoll und erfordert leistungsfähige Hardware. Für diese Arbeit wurde ein Cloud-Server verwendet, um die hohen Anforderungen für das Errechnen von insgesamt 36 kausalen Effekte (jeweils sechs Modelle für insgesamt sechs Variablen) zu decken. Ein vollständiger Durchlauf des Codes auf der verwendeten Hardware dauerte etwa 8 Stunden - trotz geringem Hyperparameter-Tuning. Ohne diese Limitationen wäre es interessant gewesen, zusätzliche Variablen aus dem Oregon Health Insurance Experiment betrachten zu können. Im Rahmen von weiteren Studien zum OHIE wurde eine Vielzahl von Variablen untersucht, die in dieser Arbeit aufgrund der technischen Limitationen im vorgegebenen Zeitraum nicht betrachtet werden konnten.

7 Schlussbetrachtung

Nachdem im Rahmen dieser Arbeit der Datensatz des Oregon Health Insurance Experiments mithilfe von Double Machine Learning auf kausale Effekte untersucht wurde, lassen sich verschiedene Punkte zusammenfassend betrachten. Zum einen lässt sich sagen, dass das OHIE einen wichtigen Beitrag zur Evaluierung der Auswirkungen von MedicAid-Programmen und öffentlichen Gesundheitssystemen auf verschiedene Faktoren liefern kann. Zum anderen lässt sich die Leitfrage, ob kausale Effekte durch das Treatment MedicAid vorliegen beantworten - ja, die Aufnahme in ein MedicAid-Programm hat kausale Effekte auf verschiedene Variablen. Inwieweit abhängige Variablen durch das Treatment beeinflusst werden, wurde in den vorigen Kapiteln ausgewertet. Damit lassen sich die in der Einführung formulierten Leitfragen teilweise beantworten:

- MedicAid hatte einen vernachlässigbar geringen positiven Einfluss auf die selbsteingeschätzte Gesundheit der Begünstigten (0,0269-0,0488)
- MedicAid hatte einen positiven kausalen Effekt auf die Anzahl der Arztbesuche (0,193)
- MedicAid hatte keinen oder einen nur sehr geringen (positiven) Effekt auf die Zufriedenheit der Teilnehmer gehabt (0,0292-0,1227)
- MedicAid hatte keinen Effekt darauf, ob alle benötigten Medikamente erhalten wurden
- MedicAid hatte einen vernachlässigbar geringen negativen Einfluss auf die Qualität der erhaltenen medizinischen Leistungen

Nicht beantworten lassen sich mit den Ergebnissen dieser Arbeit, ob Menschen ohne MedicAid-Coverage auf eigentlich notwendige medizinische Leistungen verzichten. Dennoch hat sich die Anwendung von Double Machine Learning als größtenteils erfolgreich erwiesen - es wurde jedoch auch ersichtlich, dass ein geeigneter Datensatz von hinreichender Größe notwendig ist, um signifikante Ergebnisse zu erhalten. Zudem hat sich gezeigt, dass es in der Anwendung der verschiedenen Modelle durchaus zu Unterschieden gekommen ist. Aus diesem Grund erscheint es sinnvoll, auch in zukünftigen Anwendungen von Double Machine Learning weiterhin verschiedene Modelle gegeneinander zu testen und den kausalen Effekt konservativ als ein Intervall zu schätzen. Solange nicht die Annahme besteht, dass im Datensatz des OHIE wesentliche Confounder ausgelassen wurden oder der Umgang mit Missing Values den Datensatz signifikant verzerrt hat, ist davon auszugehen, dass die Ergebnisse dieser Arbeit valide und robust sind.

8 Ausblick

Ein wesentliches Ziel dieser Arbeit, neben der Beantwortung der Leitfragen, war eine Grundlage für zukünftige Literatur und Forschung zu schaffen. Kausale Inferenz ist eine noch junge akademische Disziplin, die eine hohe Einstiegsbarriere hat. Da Statistik jedoch unerlässlich für die Forschung ist und Kausale Inferenz ebenfalls vielversprechende Werkzeuge mitbringt, hält der Autor es für wichtig, diese Werkzeuge so zugänglich wie möglich zu gestalten. Aus diesem Grund wurde im Rahmen dieser Arbeit viel Wert darauf gelegt, Prozesse transparent und nachvollziehbar zu beschreiben und - falls möglich - Entscheidungen verständlich zu begründen. Das Lesen dieser Arbeit soll den Grundstein für den Einstieg in das Thema Kausale Inferenz und Double Machine Learning, sowie das *DoubleML*-Package schaffen können, damit in Zukunft häufiger tatsächliche kausale Effekte statt Assoziationen untersucht werden. Der bereitgestellte Code kann durch den modularen Aufbau leicht abgewandelt und für andere Analysen verwendet werden. Wie im Hauptteil der Arbeit gesehen werden konnte, können Double Machine Learning und klassische statistische Methoden sich in ihren Ergebnissen unterscheiden. Aus diesem Grund wäre es zukünftig von großem Interesse, weiterhin Ergebnisse beider Ansätze auf identischen Datensätzen zu vergleichen. Ebenfalls kann es sinnvoll sein, ältere Studien von hoher akademischer Relevanz, falls möglich, unter Anwendung von Double Machine Learning erneut zu betrachten.

Anhang

Der vollständige Code zur Datenverarbeitung, Modellimplementierung und Analyse ist in der beigefügten Jupyter-Notebook-Datei auf dem elektronischen Datenträger enthalten.

Literatur

Acuna, E., & Rodriguez, C. (2004). The Treatment of Missing Values and Its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, 639–647.

Allison, P. D., et al. (2010). *Missing Data* (Bd. 200210). Sage Thousand Oaks, CA.

Angrist, J., & Imbens, G. (1995, Februar). *Identification and Estimation of Local Average Treatment Effects* (Techn. Ber.). National Bureau of Economic Research. https://doi.org/10.3386/t0118

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (n. d.). *DoubleML in Python*. https://docs.doubleml.org/stable/guide/basics.html (accessed: 01.03.2023).

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2021). DoubleML – an Object-Oriented Implementation of Double Machine Learning in R. https://doi.org/10.48550/ARXIV.2103.09603

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2022). DoubleML – an Object-Oriented Implementation of Double Machine Learning in Python. *Journal of Machine Learning Research*.

Baicker, K., Allen, H., Wright, B., Taubman, S., & Finkelstein, A. (2018). The Effect of Medicaid on Management of Depression: Evidence From the Oregon Health Insurance Experiment: the Effect of Medicaid on Management of Depression. *The Milbank Quarterly*, *96*, 29–56. https://doi.org/10.1111/1468-0009.12311

Baicker, K., & Finkelstein, A. (2014). *Oregon Health Insurance Experiment*. https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment?page=1%5C&perPage=50

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., & Finkelstein, A. N. (2013). The Oregon Experiment — Effects of Medicaid on Clinical Outcomes [PMID: 23635051]. *New England Journal of Medicine*, *368*(18), 1713–1722. https://doi.org/10.1056/NEJMsa1212321

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-parameter Optimization. *Journal of Machine Learning Research*, *13*(2).

Birkmeyer, J. D., Siewers, A. E., Finlayson, E. V., Stukel, T. A., Lucas, F. L., Batista, I., Welch, H. G., & Wennberg, D. E. (2002). Hospital Volume and Surgical Mortality in the United States. *New England Journal of Medicine*, *346*(15), 1128–1137. https://doi.org/10.1056/nejmsa012337

Bramer, M. (2007). Avoiding Overfitting of Decision Trees. *Principles of Data Mining*, 119–134.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (2001). *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Bro, R., Kjeldahl, K., Smilde, A. K., & Kiers, H. (2008). Cross-validation of Component Models: a Critical Look At Current Methods. *Analytical and Bioanalytical Chemistry*, *390*, 1241–1251.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A Method for Assessing the Quality of a Randomized Control Trial. *Controlled Clinical Trials*, *2*(1), 31–49. https://doi.org/https://doi.org/10.1016/0197-2456(81)90056-8

Chambers, R. L. (2003, März). *Analysis of Survey Data*. Wiley. https://doi.org/10.1002/0470867205

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, *107*(5), 261–65. https://doi.org/10.1257/aer.p20171038

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2016, Juli). *Double/Debiased Machine Learning for Treatment and Causal Parameters* (Papers Nr. 1608.00060). arXiv.org. https://ideas.repec.org/p/arx/papers/1608.00060.html

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*. https://doi.org/10.2139/ssrn.1819486

Coase, R. (n. d.). Quote: Ïf You Torture the Data Long Enough, It Will Confess To Anything."

Cournapeau, D. (n. d.). *sklearn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed: 22.06.2022).

Croxton, F. E., & Cowden, D. J. (1939). *Applied General Statistics.* Prentice-Hall, Inc. https://doi.org/10.1037/13608-000

Davey Smith, G. (2009). Smoking and lung cancer: causality, Cornfield and an early observational meta-analysis. *International Journal of Epidemiology*, *38*(5), 1169–1171. https://doi.org/10.1093/ije/dyp317

Dickman, S. L., Himmelstein, D. U., & Woolhandler, S. (2017). Inequality and the Health-care System in the USA. *The Lancet*, *389*(10077), 1431–1441. https://doi.org/https://doi.org/10.1016/S0140-6736(17)30398-7

Durrett, R. (2019, April). *Probability* (5. Aufl.). Cambridge University Press.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & Group, O. H. S. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year*. *The Quarterly Journal of Economics*, *127*(3), 1057–1106. https://doi.org/10.1093/qje/qjs020

Fradkin, D., & Madigan, D. (2003). Experiments with Random Projections for Machine Learning, 517–522. https://doi.org/10.1145/956750.956812

Georgii, H.-O. (2009, Juni). *Stochastik*. Walter De Gruyter. https://doi.org/10.1515/9783110215274

Granger, C. W. (1986). Statistics and Causal Inference By P. Holland. *Journal of the American Statistical Association*, *81*(396), 967–968.

Greenland, S., Pearl, J., & Robins, J. M. (1999a). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48.

Greenland, S., Pearl, J., & Robins, J. M. (1999b). Confounding and Collapsibility in Causal Inference. *Statistical Science*, *14*(1). https://doi.org/10.1214/ss/1009211805

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A Critique of the Cross-lagged Panel Model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Hawkins, D. M. (2003). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Heckman, J. (1977, März). *Sample Selection Bias As a Specification Error (with an Application To the Estimation of Labor Supply Functions)* (Techn. Ber.). National Bureau of Economic Research. https://doi.org/10.3386/w0172

Hernan, M. A., & Robins, J. M. (2023, Mai). *Causal Inference*. CRC Press.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A Practical Guide To Support Vector Classification.

Hussey, P. S., Schneider, E. C., Rudin, R. S., Fox, D. S., Lai, J., & Pollack, C. E. (2014). Continuity and the Costs of Care for Chronic Disease. *JAMA Internal Medicine*, *174*(5), 742–748.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kantarjian, H., & Rajkumar, S. V. (2015). Why Are Cancer Drugs So Expensive in the United States, and What Are the Solutions? *Mayo Clinic Proceedings*, *90*(4), 500–504. https://doi.org/https://doi.org/10.1016/j.mayocp.2015.01.014

Kohavi, R., & Wolpert, D. (1997). Bias Plus Variance Decomposition for Zero-One Loss Functions.

Malthus, T. R. (1888). *An Essay on the Principle of Population: Or, a View of Its Past and Present Effects on Human Happiness*. Reeves & Turner.

Nichols, A. (2006). *Weak Instruments: an Overview and New Techniques* (North American Stata Users' Group Meetings 2006). Stata Users Group. https://EconPapers.repec.org/RePEc:boc:asug06:3

Ott, J., Ullrich, A., & Miller, A. (2009). The Importance of Early Symptom Recognition in the Context of Early Detection and Cancer Survival. *European Journal of Cancer*, *45*(16), 2743–2748. https://doi.org/https://doi.org/10.1016/j.ejca.2009.08.009

Oxford Dictionary of English. (2010, Januar). Oxford University Press. https://doi.org/10.1093/acref/9780199571123.001.0001

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference* (2nd). Cambridge University Press.

Pearl, J. (2009b). Causal Inference in Statistics: an Overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-ss057

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and Tuning Strategies for Random Forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What To Do When Data Are Missing in Group Randomized Controlled Trials. NCEE 2009-0049. *National Center for Education Evaluation and Regional Assistance*.

Resnik, D. B., & Vorhaus, D. B. (2006). Genetic Modification and Genetic Determinism. *Philosophy, Ethics, and Humanities in Medicine*, *1*, 1–11.

Roberts, P., Priest, H., & Traynor, M. (2006). Reliability and Validity in Research. *Nursing Standard*, *20*(44), 41–45. https://doi.org/10.7748/ns2006.07.20.44.41.c6560

Romeijn, J.-W. (2022). Philosophy of Statistics. In E. N. Zalta & U. Nodelman (Hrsg.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: the Fisher Randomization Test Comment. *Journal of the American Statistical Association*, *75*(371), 591. https://doi.org/10.2307/2287653

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter Tuning and Performance Assessment of Statistical

and Machine-learning Algorithms Using Spatial Data. *Ecological Modelling*, *406*, 109–120.

SGB V. (1989).

Suits, D. B. (1957). Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association*, *52*(280), 548–551. https://doi.org/10.1080/01621459.1957.10501412

Summers, J., & Morrison, E. (2009). Principles of Healthcare Ethics. *Health Care Ethics. 2nd Ed. Sudbury: Jones and Bartlett Publishers*, 41–58.

Syrgkanis, V., & Zampetakis, M. (2020). Estimation and Inference with Trees and Forests in High Dimensions. https://doi.org/10.48550/ARXIV.2007.03210

Thompson, R. P., & Upshur, R. E. (2017, August). *Philosophy of Medicine*. Routledge. https://doi.org/10.4324/9781315159843

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Verfügbar 21. März 2023 unter http://www.jstor.org/stable/2346178

VanderWeele, T. J., & Shpitser, I. (2013). On the Definition of a Confounder. *The Annals of Statistics*, *41*(1). https://doi.org/10.1214/12-aos1058

Vladeck, B. (2003). Universal Health Insurance in the United States: Reflections on the Past, the Present, and the Future. *American Journal of Public Health*, *93*(1), 16–19. https://doi.org/10.2105/ajph.93.1.16

von Winterfeld, D., & Edwards, W. (1986, September). *Decision analysis and behavioral research*. Cambridge University Press.

Wang, L., & Alexander, C. (2016). Machine Learning in Big Data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*, 52–61. https://doi.org/10.33889/IJMEMS.2016.1.2-006

White, F. (2015). Primary Health Care and Public Health: Foundations of Universal Health Systems. *Medical Principles and Practice*, *24*(2), 103–116. https://doi.org/10.1159/000370197

Wong, J., Manderson, T., Abrahamowicz, M., Buckeridge, D., & Tamblyn, R. (2019). Can Hyperparameter Tuning Improve the Performance of a Super Learner?: a Case Study. *Epidemiology*, *30*, 1. https://doi.org/10.1097/EDE.0000000000001027

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*, *20*(7), 557–585.

Yarlett, D. (2002). Uncertainty in Causal and Counterfactual Inference.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine Learning on Big Data: Opportunities and Challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/https://doi.org/10.1016/j.neucom.2017.01.026

Abkürzungsverzeichnis

bzw. Beziehungsweise

cet. par. ceteris paribus

d. h. das heißt

E. D. eigene Darstellung

etc. et cetera

et al. et alia

IIVM Interactive IV Model

IRM Interactive Regression Model

IV Instrumentalvariable

n. d. no date

OHIE Oregon Health Insurance Experiment

OLS Ordinary Least Squares

z.B. zum Beispiel

Symbolverzeichnis

ε Störterm

ω kausaler Effekt

ϱ Zufallsfehler

ϑ Standardabweichung

ϖ kausaler Effekt