

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

From Prediction to Action: Supporting Medical Decisions with Causal Inference and Actionable AI

Leonor Alves Costa



Master's in Bioengineering

Supervisor: Vânia Guimarães, Senior Scientist Fraunhofer Portugal AICOS

Second Supervisor: Carolina Carvalho, Scientist Fraunhofer Portugal AICOS

Academic Supervisor: Moisés Santos, Assistant Professor FEUP

September 5, 2026

From Prediction to Action: Supporting Medical Decisions with Causal Inference and Actionable AI

Leonor Alves Costa

Master's in Bioengineering

Approved in oral examination by the committee:

Chair: Prof. Ana Maria Mendonça

External Examiner: Prof. Robson Bonidia

Supervisor: Prof. Moisés Santos

September 5, 2026

Abstract

Mechanical ventilation is a life-saving intervention for critically ill patients admitted to the intensive care unit (ICU), but its use is associated with multiple complications, such as ventilator-associated infections, prolonged ICU stays and increased mortality, which may be mitigated by reducing time spent on the ventilator. This procedure can often only be tolerated with appropriate sedation, achieved through the titration of propofol, yet, previous research has mostly focused on agent comparisons and sedation targets while neglecting the question of the causal effects of specific propofol infusion rate strategies on ventilatory outcomes. Moreover, the limited studies addressing causal estimation have relied on reinforcement learning, which optimizes treatment decisions instead of directly estimating causal contrasts, hindering clinical acceptance and implementation.

This dissertation applied a Causal Inference pipeline to evaluate the effects of five different propofol dosing strategies on ventilator-free days at 28 days (VFD-28) in mechanically ventilated sepsis patients, using a publicly available electronic health record dataset.

Ensemble-based Causal Discovery, comprising three distinct algorithms, was implemented to derive a causal graph, from which eleven variables that may influence VFD-28 were identified. Then, the parametric G-formula was applied, using the derived adjustment set, to estimate the causal effects of each intervention – no propofol administration, maintenance dosing (0-50 mcg/kg/min), above-maintenance dosing (50-67 mcg/kg/min), high propofol doses (>67 mcg/kg/min) and a dynamic regime following sedation depth targets – on the outcome. The analysis revealed that above-maintenance dosing showed the most positive effect on the cohort, with an estimated increase of +0.64 VFD-28 compared to observed clinical practice (95% CI 0.09, 1.17). Conversely, withholding propofol was associated with fewer ventilator-free days (-0.43 VFD-28 95% CI -0.67, -0.18). The dynamic strategy, targeting light sedation, also resulted in a negative estimated effect, despite current clinical recommendations supporting sedative titration to achieve adequate sedation depth. Furthermore, the majority of estimates remained stable across refutation tests; however, the small E-values indicate that they may be vulnerable to unmeasured confounding.

These findings suggest that modest increments to standard maintenance propofol dosing ranges may be associated with more ventilator-free days at 28 days in mechanically ventilated sepsis patients, although future research should address latent confounding. Moreover, while these results are insufficient to infer optimal propofol dosing strategies, guide changes in sedation management practices or replace clinical reasoning, they support subsequent work into interpretable approaches for treatment optimization in critically ill patients.

UN Sustainability Development Goals

The Sustainable Development Goals (SDGs) are an initiative put forward by the United Nations (UN) to guide global authorities in policy making for a more just and sustainable planet by 2030. Since its adoption in 2015, all UN member states have been tasked with addressing the seventeen interconnected goals identified as urgent global challenges, including the eradication of poverty, improving access to healthcare and education, combating climate change and fostering peace.

The present dissertation primarily contributes to two SDGs described in Table 1.

Table 1: Identification of Sustainable Development Goals and work contribution.

SDG	Target	Contribution	Performance Indicators and Metrics
3: Good health and well-being	3.8: Achieve universal health coverage, including access to quality essential healthcare services	Development of a Causal Inference pipeline to evaluate the effects of propofol sedation strategies in mechanically ventilated ICU patients with sepsis, in order to reduce ventilator dependence and improve patient outcomes. Additional considerations aim to reduce the economic burden of continuous ventilatory support on hospital infrastructures.	Estimated 0.64 (95% CI 0.09, 1.17) additional ventilator-free days at 28 days under above-maintenance dosing (50-67 mcg/kg/min), which was stable across refutation tests but sensitive to unmeasured confounding (E-value equal to 1.24), potentially improving patient outcomes and ICU resource use while reducing costs.
9: Industry, Innovation and Infrastructure	9.b: Support domestic technology development, research and innovation, including by ensuring a conducive policy and regulatory environment		

Agradecimentos

Esta dissertação é dedicada a todos aqueles que, dos campos às fábricas, e a tantas outras formas de labuta, trabalharam desde cedo para que as gerações seguintes pudessem ir mais longe. À minha família, um profundo obrigado.

A conclusão deste trabalho não teria sido possível sem o apoio e orientação de muitas pessoas, às quais gostaria de expressar a minha gratidão.

Primeiramente, à Vânia Guimarães, agradeço a confiança que depositou em mim, na minha curiosidade e na minha vontade de aprender, e por me ter dado a oportunidade de desenvolver este trabalho. À Carolina Carvalho, agradeço a ajuda inestimável na compreensão da causalidade e dos mecanismos por detrás da descoberta causal. A sua disponibilidade e apoio foram essenciais ao longo deste percurso. Agradeço ainda ao professor Moisés Santos, que sempre me orientou, encorajou e me ensinou a aceitar que, na Ciência, os resultados não são sempre aqueles que esperamos. Deixo também o meu agradecimento à Fraunhofer Portugal AICOS, por me ter permitido explorar as aplicações da inferência causal ao suporte à decisão médica e pelos recursos e condições disponibilizados para o efeito.

To my friends, some of whom have shared the most demanding parts of this degree with me, and others who are scattered across the world, thank you for lending a listening ear and kind words. And thank you for not taking it to heart when the work got overwhelming and I went MIA. You are the best and have made this process easier. To BEST Porto and to my Paw Patrol compadres, thank you for the amazing memories. You allowed me to step outside my comfort zone, learn way beyond what happens within a classroom, meet incredible people and share numerous moments of laughter and complicity. My university experience would not have been the same without you. To Bea, who has shared this journey alongside me, through the good and the bad: we made it. Thank you for always being there to listen to my rants, for the mutual support and for knowing when I needed a friend, a laugh or just someone who understood.

To Faizan, who was probably the hardest thank you to put into words, thank you for the “mechanical sheep” pep talks, for giving me perspective when I was overwhelmed, for distracting me when I needed it and above all, for the love. Aos meus irmãos, um grande obrigada por me chatearem o suficiente para me convencerem a fazer pausas para jogar Monopólio. E ao Cookie, o meu colega de trabalho felino, pela companhia e pelas pausas necessárias para fazer festinhas.

Por fim, o meu mais sincero obrigado vai para os meus pais, que sem eles nunca teria tido a oportunidade de ingressar no Ensino Superior. Ao meu pai, agradeço o incentivo e a companhia quando trabalhava pela noite fora. À minha mãe, que nunca poderia ter pedido por uma melhor, agradeço todo o apoio, todo o cuidado e toda a dedicação que me teve, mesmo nos momentos mais difíceis. Acima de tudo, agradeço-vos o amor que sempre me deram.

Estes foram anos de muita luta, muita perda e também muito ganho. Terminar agora este percurso como Mestre é, para mim, motivo de grande orgulho e gratidão, sabendo que tal não teria sido possível sem todos vós.

Leonor Costa

Contents

1	Introduction	1
1.1	Mechanical Ventilation in the ICU	1
1.2	Sedation Protocols in Mechanically Ventilated Patients	3
1.3	Propofol	5
1.4	Sepsis	6
1.5	Motivation and Objectives	6
1.6	Document Outline	7
2	Background	9
2.1	Foundations of Causality	9
2.2	Graphical Models for Causality	10
2.2.1	Assumptions	11
2.2.2	Back-door Criterion and Adjustment Set	12
2.3	Causal Discovery	13
2.3.1	Peter-Clark Algorithm	13
2.3.2	Fast Greedy Equivalence Search Algorithm	14
2.3.3	Best Order Search Structure Algorithm	15
2.4	Causal Inference	15
2.4.1	The G-formula	16
2.4.2	Assumptions	17
2.4.3	Causal Effect Estimand	17
2.5	Evaluation Metrics	18
2.5.1	E-values	18
2.5.2	Refutation Tests	18
3	Literature Review	21
3.1	Causal Inference in the ICU	21
3.2	Sedation Management in the ICU	23
3.2.1	Causal Inference Methods to Study Sedation in the ICU	25
3.3	Dissertation Positioning	27
4	Methods	29
4.1	Data	29
4.1.1	Data source	29
4.1.2	Cohort Definition	29
4.1.3	Datasets	32
4.1.4	Data Processing	35
4.2	Causal Discovery	36

4.2.1	Algorithm Selection	36
4.2.2	Background Knowledge	36
4.2.3	Ensemble Approach	37
4.2.4	Adjustment Set Identification	37
4.3	Causal Inference	38
4.3.1	Covariate Specification	38
4.3.2	Outcome Model	38
4.3.3	Model Assessment	39
4.3.4	Treatment Strategies	39
4.4	Robustness and Sensitivity Analysis	40
5	Results	41
5.1	Cohort Characteristics	41
5.2	Causal Discovery Results	45
5.3	Causal Inference Results	46
5.4	E-values and Refutation Tests	48
5.5	Discussion	49
6	Conclusion	55
6.1	Future Work	56
	References	57

List of Figures

1.1	Propofol structure and mechanism of action.	5
2.1	Fundamental causal structures.	11
2.2	Illustration of the PC algorithm.	14
4.1	Study cohort flowchart.	31
4.2	Temporal tiers and forbidden edges identified in the study dataset.	37
5.1	Subgraph of the consensus causal DAG illustrating the adjustment set for the relationship between propofol rate and VFD-28.	45
5.2	Comparison of simulated covariate trajectories and non-parametric means across five epochs, for nine key time-varying covariates.	47

List of Tables

1	Identification of Sustainable Development Goals and work contribution.	iii
1.1	Richmond Agitation-Sedation Scale.	4
3.1	Causal approaches to estimate the effects of different sedatives and analgesics in the ICU.	26
4.1	Aggregation map for baseline covariate extraction.	33
4.2	Static propofol infusion rate strategies.	39
5.1	Summary of baseline characteristics and outcomes of the study cohort.	42
5.2	Standardized differences across propofol dose categories	44
5.3	Estimated ventilator-free days at 28 days under natural course and hypothetical propofol dosing interventions using the parametric G-formula.	46
5.4	Sensitivity analysis including E-values and refutation test results by intervention.	49

Abbreviations

AI	Artificial Intelligence
APS III	Acute Physiology Score III
ARDS	Acute Respiratory Distress Syndrome
ATE	Average Treatment Effects
BOSS	Best Order Search Structure
CCI	Charlson Comorbidity Index
CI	Confidence Interval
CPDAG	Completed Partially Directed Acyclic Graph
DAG	Directed Acyclic Graph
ECMO	Extracorporeal Membrane Oxygenation
EHR	Electronic Health Records
fGES	Fast Greedy Equivalence Search
FiO ₂	Fraction of Inspired Oxygen
GCS	Glasgow Coma Score
ICU	Intensive Care Unit
IPTW	Inverse-Probability-of-Treatment Weighting
MAP	Mean Arterial Pressure
MEC	Markov Equivalence Class
MIMIC-IV	Medical Information Mart for Intensive Care IV
MV	Mechanical Ventilation
NICE	Non-Iterative Conditional Expectation
PADIS	Clinical Practice Guidelines for the Prevention and Management of Pain, Agitation/Sedation, Delirium, Immobility, and Sleep Disruption in Adult Patients
PC	Peter-Clark
PEEP	Positive End-Expiratory Pressure
PRIS	Propofol-Related Infusion Syndrome
PSM	Propensity Score Matching
RASS	Richmond Agitation-Sedation Scale
RCTs	Randomised Controlled Trials
RL	Reinforcement Learning
SATs	Spontaneous Awakening Trials
SBTs	Spontaneous Breathing Trials
SCCM	Society of Critical Care Medicine
SMDs	Standardized Mean Differences
SOFA	Sequential Organ Failure Assessment
SpO ₂	Peripheral Capillary Oxygen Saturation
TMLE	Targeted Maximum Likelihood Estimation
VAP	Ventilator-Associated Pneumonia
VFD-28	Ventilator-Free Days at 28 days

Chapter 1

Introduction

“First, do no harm”, is a popular expression that captures the medical ethical principle of non-maleficence [1]. Yet, in the intensive care unit, life-saving procedures are often inherently harmful. Mechanical ventilation is one such intervention that may be essential to sustain critically ill patients, but it requires sedation to be tolerated. Sedation, however, requires a fine balance between managing patient discomfort and the risk of suppressing consciousness and prolonging intubation periods [2]. Still, finding this equilibrium and defining an optimal sedation strategy remains an open question in critical care.

1.1 Mechanical Ventilation in the ICU

Mechanical ventilation (MV) is the most common form of respiratory support in the intensive care unit (ICU), required by 20-40% of ICU admissions in the United States [3] and present in 39-49% of ICU patients at any given time [4]. Notably, most ICU patients are started on mechanical ventilation within the first 24 hours of admission [4] and while the majority is successfully stabilized and weaned from the ventilator within 3 to 4 days, 1 in 20 MV patients necessitates prolonged mechanical ventilation, consistent with continuous respiratory support for over 21 days [5]. Furthermore, the economic burden of MV is substantial and accounts for most of ICU expenses worldwide [6][7]. The approximate daily costs for each ventilated patient are 59% superior to a non-ventilated patient [8], totalling more than 1000€/day in France and Germany [7] and an additional 1522\$ per ICU day in the United States [6].

The primary goal of mechanical ventilation is to decrease the work of breathing while maintaining adequate pulmonary gas exchange, by supporting or substituting the function of the respiratory muscles [9]. During spontaneous breathing, the inspiratory muscles, such as the diaphragm,

generate a negative pleural pressure ¹, which creates a positive transpulmonary gradient that allows the lungs to expand. In MV, this gradient is dependent on the delivery of a positive airway pressure that drives air into the lungs and increases alveolar pressure, with passive expansion of the chest wall [10]. On a ventilator, one may adjust airway pressure, flow and timing to match the patient's respiratory mechanics, and ventilation modes differentiate between fully controlled and assisted breathing, where the latter shares the work of breathing between ventilator and patient [11].

This procedure is predominantly indicated for airway protection in patients with decreased consciousness (e.g., head trauma, stroke, drug overdose, anesthesia), hypercapnic respiratory failure owing to airway, chest wall, or neuromuscular disease, hypoxemic respiratory failure or circulatory failure [11][12].

Despite its supportive role, mechanical ventilation is associated with severe complications. For instance, disproportionate ventilatory assistance can lead to acquired limb and diaphragmatic weakness, which hinder weaning [13][14]. Furthermore, patient-ventilator dyssynchrony, a common mismatch between patient breathing and ventilator timing, is associated with prolonged ventilation, longer ICU stays, and increased mortality [15][16][11]. Mechanically ventilated patients are also at risk for aspiration, atelectasis (i.e. lung collapse), pulmonary edema and delirium (up to 80% of patients [17]) as well as ventilator-associated conditions, categorized by objective criteria into infection-related ventilator-associated conditions and further into possible or probable ventilator-associated pneumonia (VAP) [18]. In the ICU, VAP, an infection of the lung parenchyma developing after 48 hours of intubation and mechanical ventilation, affects as many as 36% of critically ill patients, making it the most prevalent hospital-acquired infection in this population [19]. VAP is associated with increased mortality (mortality rates between 24-76%), septic shock, acute respiratory distress syndrome (ARDS), prolonged hospitalization (adding 4-9 days) and increased healthcare costs. It also contributes to multi-drug resistance due to extensive antibiotic use. Expediting ventilatory liberation reduces the risk of ventilator-associated pneumonia [19].

Ventilator-induced lung injury is another example of MV complications and refers to biological damage to the lung directly caused by this intervention via alveolar over-distention (volutrauma), excessive trans-alveolar pressure (barotrauma), high shear forces from repetitive opening and collapse of alveoli (atelectotrauma) and/or inflammation (biotrauma) [20]. Moreover, positive-pressure mechanical ventilation is linked to haemodynamic instability. The increase in intrathoracic pressure reduces venous return, ventricular preload and left ventricular afterload, lowering cardiac output (i.e. hypotension). Conversely, it increases right ventricular afterload. Thus, the haemodynamic effects of MV may be inconsequential to a healthy individual yet be significant in critically ill patients [21].

Mortality rates among patients receiving ventilatory support vary according to ICU setting and available resources. International cohorts report mortality rates ranging from 28% to 51%, increasing to 88.5% in low-resource areas. Research into mortality risk factors identify sepsis, a Glasgow

¹In the space between the lungs and the chest wall (pleural space), the pressure is around -4 mmHg at rest, approximating -6 mmHg during inspiration.

coma score (GCS) below 8, multiple organ dysfunction syndrome and the use of vasopressors as indicators [22]. Additional factors include age, sex, illness severity scores such as sequential organ failure assessment (SOFA) score and acute physiology and chronic health evaluation II (APACHE II) score, serum albumin level, need for renal replacement therapy, comorbidity burden, duration of mechanical ventilation and sedation use [22].

Given these clinical consequences associated with MV, it is important to evaluate patient outcomes that capture both the duration of respiratory support and mortality. Ventilator-free days at 28 days (VFD-28) is a commonly used composite outcome in critical care that integrates duration of MV and patient survival within a 28-day period [23].

Beyond the aforementioned physiological risks, the nature of this procedure can also be a source of pain and anxiety for critically ill patients, motivating the administration of sedatives [24].

1.2 Sedation Protocols in Mechanically Ventilated Patients

Sedation in mechanically ventilated patients aims to ease the psychological stress, physical discomfort and painful experiences related to routine nursing processes, such as endotracheal suctioning and patient repositioning, as well as decrease the work of breathing [25][24]. Indeed, sedation has been shown to reduce oxygen consumption, autonomic hyperactivity and self-extubation episodes, improving tolerance to distressing interventions [26][24][27].

Early sedation protocols in the intensive care unit were adopted from anaesthesiology practices which favoured deeply sedated mechanically ventilated patients to control the patient's stress response and support synchrony between the patient and the ventilator [28]. However, current critical care sedation recommendations, as formalized in the Clinical Practice Guidelines for the Prevention and Management of Pain, Agitation/Sedation, Delirium, Immobility, and Sleep Disruption in Adult Patients in the ICU (PADIS), encourage the minimum level of sedation compatible with patient comfort and safety [29] [2]. This shift originates from an increasing interest in sedation research over the years. Nevertheless, while randomised controlled trials (RCTs) have mostly focused on sedative agents rather than sedation strategies and have shown limited benefit of either alternative, observational studies have described stronger associations and advantages of following a light sedation approach [30][31]. In particular, cohort studies have reported worse outcomes linked to deeper sedation, including higher hospital mortality in COVID-19 patients, fewer ventilator-free days and longer ICU and hospital stays, compared to lighter sedation [32][33][34][35]. Collectively, these findings have formed the basis for the 2018 PADIS guidelines and the 2025 Society of Critical Care Medicine (SCCM) Focused Update which support a proactive, goal-directed sedation strategy that aligns early light sedation with continuous and structured monitoring via sedation depth scales [2][36].

Table 1.1: Richmond Agitation-Sedation Scale. Adapted from Sessler et. al (2002) [37].

Score	Term	Description
+4	Combative	Overtly combative or violent; immediate danger to staff
+3	Very agitated	Pulls on or removes tube(s) or catheter(s) or has aggressive behavior toward staff
+2	Agitated	Frequent nonpurposeful movement or patient–ventilator dyssynchrony
+1	Restless	Anxious or apprehensive but movements not aggressive or vigorous
0	Alert and calm	
-1	Drowsy	Not fully alert, but has sustained (more than 10 seconds) awakening, with eye contact, to voice
-2	Light sedation	Briefly (less than 10 seconds) awakens with eye contact to voice
-3	Moderate sedation	Any movement (but no eye contact) to voice
-4	Deep sedation	No response to voice, but any movement to physical stimulation
-5	Unarousable	No response to voice or physical stimulation

Procedure

1. Observe patient. Is patient alert and calm (score 0)? Does patient have behavior consistent with restlessness or agitation (score +1 to +4)?
2. If patient is not alert, in a loud voice state patient's name and direct patient to open eyes and look at speaker. Repeat if necessary.
 - Sustained eye contact > 10 seconds (score -1)
 - Eye contact < 10 seconds (score -2)
 - Movement to voice without eye contact (score -3)
3. If patient has no response to voice, physically stimulate patient (shake shoulder, then rub sternum if needed).
 - Movement to physical stimulation (score -4)
 - No response to voice of physical stimulation (score -5)

The Richmond Agitation-Sedation Scale (RASS), for instance, is one such highly reliable and well-validated monitoring tool used to assess patient depth of sedation [37]. It characterizes a patient's state within a 10-level scale, ranging from -5, representative of unarousable coma to +4, indicating combative agitation. A lightly sedated patient would lie within the RASS interval of -2 to +1, suggesting a calm, easily roused to voice and comfortable patient under minimal analgesia. Table 1.1 provides an overview of the Richmond Agitation-Sedation Scale, along with instructions for its bedside use.

The recommendations also encourage an analgesia-first approach to treat pain prior to sedative escalation and the integration of daily spontaneous awakening trials (SATs) and spontaneous breathing trials (SBTs) to minimize sedation and promote earlier weaning from mechanical ventilation. The 2025 update expands on anxiety management and conditionally endorses dexmedetomidine over benzodiazepines and propofol for light sedation and delirium incidence reduction.

Nevertheless, propofol remains an appropriate and commonly used sedative in the ICU [38] and RCTs with septic patients have not shown mortality benefit for dexmedetomidine over propofol [39], reinforcing its relevance to this dissertation.

1.3 Propofol

Propofol (2,6-diisopropylphenol) is a lipophilic sedative-hypnotic, administered intravenously, that induces a sedated state by crossing the blood-brain barrier and binding to the postsynaptic γ -aminobutyric acid type A (GABA-A) receptor. This interaction potentiates GABA-mediated chloride influx, which hyperpolarizes the neuronal membrane, effectively reducing neuronal firing (Figure 1.1) [40][38].

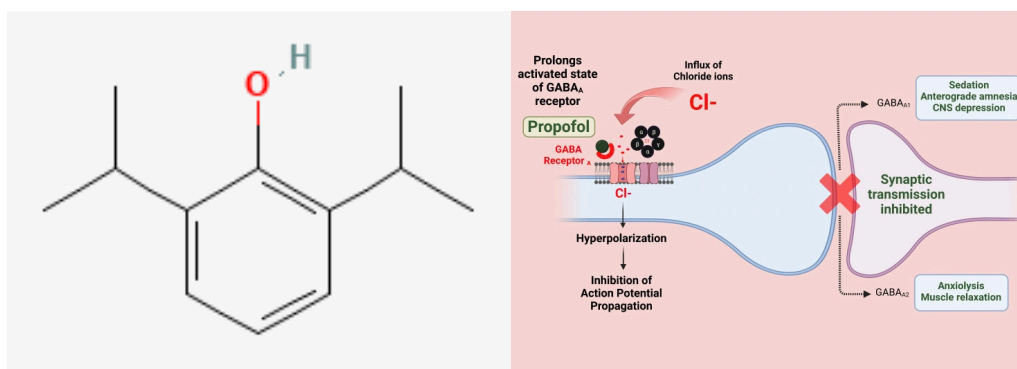


Figure 1.1: Propofol structure and mechanism of action. Left: Chemical structure of propofol (2,6-diisopropylphenol), retrieved from PubChem (CID: 4943). Right: Mechanism of action of propofol, reproduced from Paramsothy et al. 2023 [41]

Propofol's chemical properties enable a rapid onset of action (as quick as 30 seconds) and its sedative effect dissipates quickly, after short-term use, due to rapid redistribution and high hepatic and extrahepatic clearance. Despite an elimination half-life of 3-12 hours which may increase 50 ± 18.6 hours for long-term use, patients often regain consciousness within 10–30 minutes of drug discontinuation without displaying lingering sedative effects [42][43]. This fact makes propofol one of the preferred agents in sedation protocols that require frequent neurological assessment, routine SATs, and minimization of delirium episodes [26][43]. However, propofol can also be used to maintain prolonged sedation by administering continuous intravenous infusions (5-50 mcg/kg/min) or intermittent boluses (5 mcg/kg/min over 5 minutes) [40][43]. Though, it is not recommended to sustain prolonged (>48 hrs) infusion of high propofol doses (>67 mcg/kg/min), as it increments the risk for propofol-related infusion syndrome (PRIS), a condition associated with cardiorespiratory depression, severe metabolic acidosis [44] and a mortality rate as high as 33% [43].

The use of propofol has additional detrimental consequences. Pain upon injection is the most

commonly reported adverse effect, with cardiopulmonary alterations, such as bradycardia, hypotension, hypoventilation and apnoea also being frequently observed [38][43]. Moreover, long-term or high-dose use of propofol may lead to the development of hypertriglyceridemia [43].

The impact these consequences have among ICU patients depends on each individual's underlying physiological compromise [44]. As such, septic patients represent a particularly challenging subgroup.

1.4 Sepsis

According to the sepsis-3 definition, the clinical syndrome of sepsis refers to “a life-threatening organ dysfunction due to a dysregulated host response to infection” [45], leading to altered patient physiology, including drug metabolism [26]. In the ICU, patients suspect of infection are diagnosed with sepsis via their presented SOFA score or a quicker bedside table test denoted qSOFA, which is based on systolic blood pressure and respiratory rate, among other variables, and does not require laboratory testing [45]. A sepsis diagnosis is often aligned with increased haemodynamic instability, enhanced capillary permeability and compromised tissue perfusion [26][24]. In turn, sedative agent distribution may be affected by fluid resuscitation reducing propofol's effective blood concentration and possibly requiring higher doses to obtain a clinical effect. Contrarily, the lower cardiac output may be responsible for lower sedative distribution by limiting peripheral blood flow. Concurrent treatments, such as renal replacement therapy and extracorporeal membrane oxygenation (ECMO) contribute to the reduction of drug concentration by increasing its volume of distribution and through circuit sequestration. Additionally, as a high albumin binding drug, with approximately 98% bound in non-septic patients, high free propofol concentrations in the plasma, resulting from hypoalbuminemia associated with sepsis, may worsen hypotension. Furthermore, septic shock patients, a subset with escalated illness severity and mortality rates, may present impaired liver function, leading to potential drug dose accumulation and over-sedation [26].

1.5 Motivation and Objectives

Despite sedation being a central component of pain management and routinely administered in the ICU, the quality of evidence guiding sedation practices is relatively limited and inconsistent [30]. The prevailing understanding is that inadequate sedation protocols are consistently associated with worse patient outcomes, namely prolonged mechanical ventilation, longer ICU and hospitals stays, higher delirium incidence and increased mortality rates. This issue is particularly relevant in mechanically ventilated patients, which require sedative infusions to tolerate and enable synchronous respiratory support.

Propofol has historically been the agent of choice due to its rapid action onset and clearance, but optimizing its dosing remains a difficult endeavour, particularly in critically ill sepsis

patients, where altered physiology and haemodynamic instability affect dose-response relationships. Therefore, an optimal sedation strategy would be patient-specific and aimed at the lightest effective level, reducing complications and resource use, while maintaining patient comfort.

In this context, decision support algorithms may help physicians maintain appropriate sedation levels throughout a patient's stay in critical care. Causal inference methods, applied to observational ICU data, enable the simulation of alternative sedation strategies and the estimation of the effect of each approach on patient outcomes, allowing for the evaluation of treatment decisions.

Hence, the overarching objective of this Master Dissertation is to develop and evaluate a Causal Inference pipeline for estimating the causal effect of propofol dosing strategies on VFD-28 in mechanically ventilated sepsis patients. To fulfill this aim, the study focuses on three specific goals:

1. Construct a longitudinal ICU dataset of critically ill patients that captures their evolving condition under sedation management;
2. Identify a plausible set of variables that influence sedation decisions and may bias observational estimates;
3. Estimate alternative VFD-28 outcomes under diverging yet clinically grounded propofol dosing strategies and assess their robustness.

By pursuing these objectives, this work intends to illustrate both the potential and limitations of Causal Inference from observational data in studying ICU sedation management. In the process, it contributes to the growing evidence on propofol dosing in mechanically ventilated patients and applications of causal frameworks to healthcare tasks, with the long-term goal of enabling causally validated and personalised sedation protocols to improve patient outcomes.

1.6 Document Outline

The remainder of this dissertation is structured into six chapters as follows:

Chapter 2: introduces causality concepts and explores the ideas behind Causal Discovery and Causal Inference, which are foundational to the methods employed in this study.

Chapter 3: synthesizes prior contributions and current trends encompassing the evolution from predictive to causal modelling in healthcare, the clinical evidence supporting sedation guidelines for mechanically ventilated patients and the application of Causal Inference methods in critical care.

Chapter 4: describes the selection of the patient cohort and dataset preparation for downstream application in the identification and removal of confounding elements and posterior estimation of the causal effects of propofol on VFD-28.

Chapter 5: provides a comprehensive evaluation and interpretation of the results within the context of current Causal Inference research and clinical evidence.

Chapter 6: concludes this dissertation by summarizing the primary findings and contributions of the study, while acknowledging limitations and proposing future research directions.

Chapter 2

Background

Chapter 2 introduces the theoretical foundations of the methods applied in this dissertation. It covers core concepts of causality, the graphical models that represent causal structure and the algorithms that can recover this structure from observational data. Moreover, it explores the parametric G-formula as a method to approximate causal effects under hypothetical interventions and it defines the estimands used to quantify these effects.

2.1 Foundations of Causality

The current data science literature has established the efficacy of statistical and machine learning algorithms for pattern recognition, including predicting, classifying and finding associations in observational data. However, in science and healthcare, most questions are not predictive in nature but rather causal [46][47]. That is, the focus is not on knowing “which patients have fewer ventilator-free days?” but “what would happen to these patients if the sedation protocol were changed?”. Hernán et al. (2019) [48] underline the importance of not conflating both questions, as only causal analyses support decision-making by answering “what if” scenarios. Furthermore, these authors formalise this distinction through a taxonomy of data science tasks, identifying description, prediction and Causal Inference as separate analytical goals.

Similarly, Pearl (2018) [49] differentiated between types of knowledge required to answer each question by mapping relationships between variables in a hierarchical three-rung framework known as the Ladder of Causation. The first rung, association, addresses statistical relationships in data, such as determining the probability of an outcome Y given the observation of X , $P(Y | X)$. This is the sort of task performed using standard predictive methods. The second rung, concerns questions only answered through experimentation. It asks what happens to an outcome Y if X was forced to take a specific value: $P(Y | do(X = x))$. The do-operator formalises the difference between *observing* and *doing* and represents an active intervention (or treatment). Finally, the third

rung focuses on alternative scenarios (counterfactuals), corresponding to non-observed hypothetical outcomes. It aims to answer what would have been outcome Y under a different course of action. Only the second and third rungs correspond to Hernán’s definition of Causal Inference.

Naturally, observational data, such as the one contained in electronic health records (EHR), contains information about each patient’s true treatment trajectory and ultimate outcome. It does not reflect alternative outcomes under distinct treatment strategies. However, under certain assumptions explored in this chapter, Causal Inference overcomes this limitation and enables the estimation of potential outcomes under diverging interventions and the comparison of their causal effects [48][47]. As such, it allows clinically relevant questions such as “what is the propofol dosing approach that maximizes VFD-28 in septic patients?” to be addressed.

Complete causal analyses comprise two complementary steps: Causal Discovery, a process through which the underlying causal structure of the data is identified and the cause-effect relationships between the variables are uncovered; and the subsequent Causal Inference component where causal effects are quantified [50].

2.2 Graphical Models for Causality

Understanding causal structure is not always trivial; therefore, graphical models are common representations that enhance comprehension by making causal relationships explicit [51].

Directed causal graphs, $\mathcal{G} = (V, E)$, are a family of probabilistic models where V is a set of nodes corresponding to study variables, spanning treatment, outcome and additional observed and unobserved covariates, and E is a set of directed edges which encode causal dependencies between nodes. Edges are represented by arrows connecting two variables, such that an arrow from variable A to variable B , with the arrowhead on B , posits that A is the direct cause of B . When the causal model prohibits causal cycles so that no variable can cause itself through any chain of directed edges, it is referred to as directed acyclic graph (DAG). Within a DAG, one can find three fundamental path structures that represent statistical associations between variables (Figure 2.1) [52].

A chain $A \rightarrow B \rightarrow C$ denotes a mediation structure that implies that A will influence C through an intermediate variable B . Conditioning on the mediator B blocks the association along the path from A to C . A fork ($A \leftarrow B \rightarrow C$), on the other hand, is a classic example of confounding. A confounder is any variable that simultaneously influences treatment and outcome, creating a “distortion” of the treatment-outcome relationship [53]. In this context, B is a common cause of A and C which creates a spurious association between these two variables that does not reflect a causal link. Conditioning on B blocks the back-door path and removes this confounded relationship. Finally, a collider (or v-structure), $A \rightarrow B \leftarrow C$, is a path structure that illustrates how a variable can be caused by two other variables at once. Unlike the previous structures, a collider is blocked by default. Therefore, conditioning on B would open a path and create a misleading correlation between A and C [52].

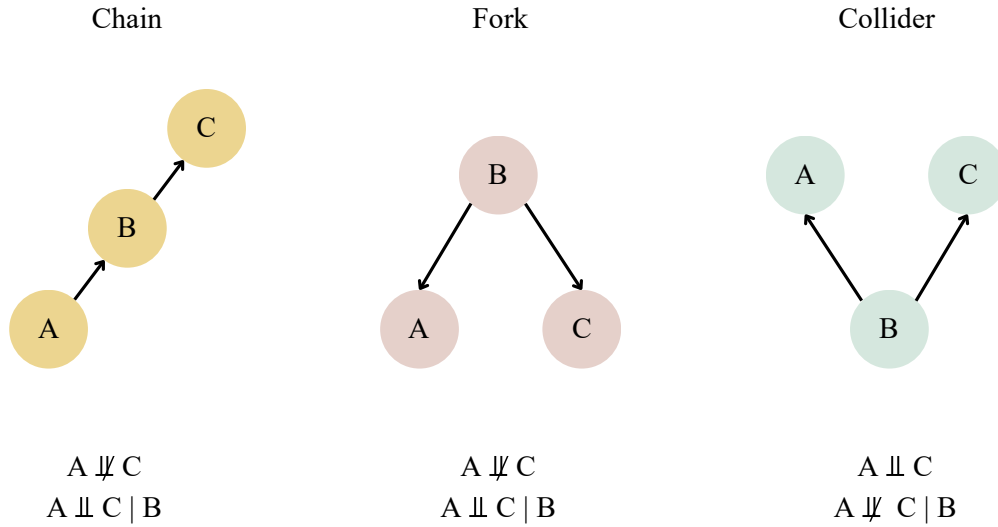


Figure 2.1: Fundamental causal structures. By the d-separation criteria, and for the chain and fork example, A and C are conditionally independent given B. However, in the collider example, A and C are independent but become conditionally dependent given B.

Despite their differences, these path structures are connected by the concept of d-separation, short for dependence separation, introduced by Pearl (1988) [54]. D-separation is a graphical criterion to determine if two variables are conditionally independent given a set of conditioning variables Z , referred to as the conditioning set.

Formally, a path between variables A and C is blocked by Z if:

1. the path contains a chain ($A \rightarrow B \rightarrow C$) or fork ($A \leftarrow B \rightarrow C$) such that the middle node $B \in Z$, or
2. the path contains a collider ($A \rightarrow B \leftarrow C$) such that $B \notin Z$ and no descendant of B is in Z.

Variables A and C are d-separated given Z if and only if Z blocks every path between them. In that case, the graph implies conditional independence, that is, $A \perp\!\!\!\perp C \mid Z$.

2.2.1 Assumptions

Although d-separation is a useful graphical criterion to identify conditional independence relations implied by a causal graph, it does not ensure that these relationships are reflected in the data. Therefore, to connect graphical structure and the underlying probability distribution, Causal Discovery relies on additional assumptions [55].

The **Causal Sufficiency Assumption** describes the absence of hidden confounding. It states that every common cause of the observed variables is also observed and included in V [56]. In other words, it assumes that all relevant data is accounted for and there are no latent confounders

that may influence the study variables. If this assumption is violated, (i.e. if there are latent confounders), the causal graph may infer fake direct causal relationships between variables which can be explained by an unmeasured confounder. Violations of the causal sufficiency assumption are particularly concerning in medical data, where unobserved variables are common and cannot be controlled for, which in turn may bias results and lead to incorrect interpretations.

The **Causal Markov Condition** states that d-separation in the graph implies conditional independence in the data distribution. The **Faithfulness Assumption** is its reverse: if the data shows conditional independence, the graph must imply d-separation. Formally, if $\mathcal{G} = (V, E)$ is a causal graph over random variables X indexed by V with joint probability distribution P , then for any disjoint subsets $A, B, C \subseteq V$ [57]:

$$A \perp\!\!\!\perp B \mid C[\mathcal{G}] \Rightarrow A \perp\!\!\!\perp B \mid C[P] \quad (2.1)$$

$$A \perp\!\!\!\perp B \mid C[P] \Rightarrow A \perp\!\!\!\perp B \mid C[\mathcal{G}] \quad (2.2)$$

Taken together, these assumptions ensure a perfect correspondence between graph structure and data patterns, indicating an absence of hidden relationships that the graph misses or spurious independences the graph fails to capture. A violation of any of these assumptions breaks this correspondence and compromises the reliability of the graph as a causal model.

The d-separation criterion also motivates the concept of a Markov Equivalence Class (MEC), which refers to a set of directed acyclic graphs that encode the same conditional independence relations [57]. Given a MEC, single DAG selection can only be achieved with additional information beyond the one obtained from observational data, such as expert knowledge [58].

In short, Causal Discovery is a process of selecting the simplest valid DAG that contains the data generating distribution, under the causal assumptions.

2.2.2 Back-door Criterion and Adjustment Set

The back-door criterion, formalised by Pearl [59], is a graphical rule applied directly to a causal graph to identify a set of variables $Z \subseteq V$ sufficient for adjustment, so that the observational estimates recover the causal effects of an intervention ($do(X = x)$) on an outcome.

A back-door path between a treatment variable X and an outcome Y is any non-causal path with an arrow pointing into X that can create a spurious association and confounding if left unblocked. A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X, Y) in a DAG $\mathcal{G}$ if [60] [61]:

1. Z blocks all back-door paths between X and Y ; and
2. Z contains no descendant of X .

The first condition guarantees that all confounding variables are blocked and the second condition ensures that no intermediate variable is adjusted for, which would block the causal path of interest and result in a bias estimate.

The adjustment set Z , determined through the back-door criterion, is a sufficient set of variables for the treatment-outcome pair (X, Y) that makes the distribution $P(Y \mid do(X = x))$ identifiable from observational data, which means that the causal effect of X on Y can be estimated based on observed probabilities (Equation 2.3) [59] [60].

$$P(y \mid do(x)) = \sum_Z P(y \mid x, z) P(z) \quad (2.3)$$

2.3 Causal Discovery

Causal Discovery methods can be generally divided into two categories: constraint-based methods and score-based methods [55]. Constraint-based algorithms leverage conditional independence tests to extract causal structure from observational data. If two variables are found to be conditionally independent given some set Z , the algorithm understands that they are d-separated by Z in the causal graph and uses this information to prune and orient edges. On the other hand, score-based methods do not test independence directly, but rather establish a scoring function, such as the degenerate Gaussian score [62], that measures the fit between a candidate graph and the observed data. These methods span the space of possible causal graphs and output the structure that maximises the chosen score. Of the three algorithms explored in the present dissertation, the Peter-Clark (PC) algorithm is constraint-based, and both Fast Greedy Equivalence Search (fGES) algorithm and Best Order Search Structure (BOSS) algorithm are score-based. The next sections expand on each of these methods.

2.3.1 Peter-Clark Algorithm

The Peter-Clark algorithm, named after its creators Peter Spirtes and Clark Glymour, is one of the first proposed methods for Causal Discovery tasks and shown to be consistent assuming independent and identically distributed data and absence of hidden confounding [55]. It takes as input a dataset of observed variables and outputs a Completed Partially Directed Acyclic Graph (CPDAG), which represents the MEC of the true causal graph.

The PC algorithm consists of two sequential phases: first, identifying the undirected skeleton of a causal graph, followed by edge direction and orientation (Figure 2.2).

The algorithm starts with a fully connected, undirected graph. Then, edges are systematically removed based on conditional independence tests, such that, for each pair of variables X and Y , the algorithm searches for a conditioning set Z , among the neighbouring variables to X or Y , in the current graph, where $X \perp\!\!\!\perp Y \mid Z$. If Z is found, then the edge between X and Y is removed and Z is now recorded as the separating set for the $\{X, Y\}$ variable pair. This process is repeated for

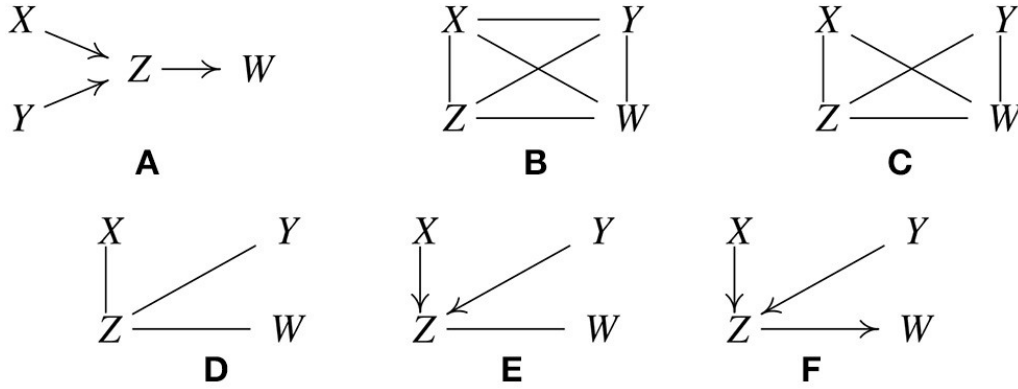


Figure 2.2: Illustration of how the PC algorithm works. (A) Original true causal graph. (B) PC starts with a fully-connected undirected graph. (C) The $X - Y$ edge is removed because $X \perp\!\!\!\perp Y$. (D) The $X - W$ and $Y - W$ edges are removed because $X \perp\!\!\!\perp W \mid Z$ and $Y \perp\!\!\!\perp W \mid Z$. (E) After finding v-structures. (F) After orientation propagation. Reproduced from Glymour (2019) [55].

growing conditioning sets until no more edges can be removed. In the example in Figure 2.2 (D), conditional independence dictates the removal of edges $X - W$ and $Y - W$.

In the edge orientation phase, the algorithm searches triple variable structures, for instance $A - B - C$, where A and C are not adjacent nodes, but both A and B and B and C are adjacent variables respectively. The edges are then oriented as $A \rightarrow B \leftarrow C$ if B was not in the separating set of A and C . This is, in essence, collider detection. In the example shown in Figure 2.2 (E), this corresponds to orienting $X - Z - Y$ as $X \rightarrow Z \leftarrow Y$, where Z , in this case, is a node in the graph and not part of the separating set of X and Y .

Following this step, the remaining edge orientations are propagated using logical rules, known as "Meek orientation rules". For example, if $A \rightarrow B - C$, and A and C are not adjacent, the edge can be oriented as $B \rightarrow C$, since orienting it as $C \rightarrow B$ would create a new collider at B . In the example shown in Figure 2.2 (F), this corresponds to orienting $X \rightarrow Z - W$ as $X \rightarrow Z \rightarrow W$. It is possible that some edges cannot be oriented without violating Meek's rules, thus they remain undirected in the output [55][63].

2.3.2 Fast Greedy Equivalence Search Algorithm

Unlike the previous method, the Greedy Equivalence Search (GES) algorithm [64] [55] starts with an empty graph and builds the causal structure through a two-phase greedy search over the space of Markov equivalence classes. Meek's Conjecture, proven by Chickering (2002) [64], is the fundamental theoretical result underlying this algorithm. It guarantees that the space of equivalence classes is fully connected, such that starting from any class, one can reach any other class by a finite sequence of edge additions and reversals. Therefore, in the limit of large samples, a greedy search over the space of MEC will identify the true equivalence class without having to evaluate every possible DAG.

In the forward phase of the algorithm, edges are added at each step based on whether its addition improves the scoring function, such as a quasi-Bayesian score like the Bayesian information

criterion, or degenerate Gaussian or even the Z-score of a statistical hypothesis test. If affirmative, the edge is drawn and the process is repeated iteratively until no further improvements are possible. In the backward phase, edges are removed by the same logic, such that only edges whose removal improves the chosen evaluation criterion are excluded. Once no additional alterations improve the score, the algorithm outputs a CPDAG. [55]

Fast greedy equivalence search (fGES) is a variant of the GES algorithm that incorporates several modifications that increase computational efficiency and scalability. The algorithm leverages the decomposability of scoring functions to cache local scores and avoid redundant computations, which, along with the fact that it is also parallelizable, enhances the search speed. Furthermore, fGES allows for higher penalty terms in scoring functions, leading to sparser graphs and reduced runtime. However, while this may be beneficial in high-dimensional settings like ICU data, this trade-off depends on the confounding structure, since excessive penalization can remove true causal edges and compromise result validity. For improved scalability, fGES may optionally include a limited faithfulness assumption to prune unlikely edges early in the search process, reducing the search space for high-dimensional datasets [65].

2.3.3 Best Order Search Structure Algorithm

The Best Order Search Structure (BOSS) algorithm [57] is a special case where the two-step causal learning process is a combination of a permutation-based search strategy and a score-based evaluation phase; the latter being analogous to the second stage in the GES algorithm. It is based on causal ordering, ensuring that no variable precedes any of its causes, and equates finding the best ordering to determining the best DAG.

Starting from an initial permutation $\pi = \langle a, b, c, \dots \rangle$, BOSS greedily constructs and refines variable ordering by iteratively repositioning each variable in the position that maximizes a penalized likelihood score. That is, for each variable $v \in V$ and index $i \leq |V|$, BOSS takes π and allows two operations: $\pi.index(v)$, which returns the position of v in the permutation, and $\pi.move(v, i)$, which creates a new permutation by removing v from its current position in π and reinserting it at position i . This process is repeated until no more improvements can be made.

Computational efficiency in this algorithm is owed to the Grow-Shrink Trees data structure which caches intermediate score results avoiding recomputation of variable subsets that remain unchanged between iterations. Furthermore, BOSS is highly parallelizable, which increases efficiency and suitability for large-scale datasets.

2.4 Causal Inference

Once a causal graph has been derived and an adjustment set identified, the causal effect of a given treatment on a chosen outcome can be estimated. This estimation relies on defining an *estimand* (or target causal quantity), selecting an appropriate *estimator*, and computing the corresponding *estimate* [50].

To that end, the parametric G-formula, an estimator under Robin's [66] generalized methods (G-methods) that can handle both time-varying covariates and treatment, was the chosen approach.

2.4.1 The G-formula

In the context of longitudinal studies, causal effect estimation must account for time-varying confounding, since treatment assignment at each time-point may depend on the history of relevant observed covariates and treatments [67]. In clinical settings, this aims to reflect the sequential-decision making that physicians face when determining a patient's treatment at each evaluation time point based on their history of clinical measurements and responses to previous treatments.

Let $j = 0, 1, \dots, J$ be the time points of follow-up, A_j be the treatment at epoch j and $Y = \text{VFD-28}$ the outcome measured at the end of the observation window. Additionally, let L_j denote the vector of time-varying confounders at epoch j and $\bar{A}_j = (A_0, \dots, A_j)$ and $\bar{L}_j = (L_0, \dots, L_j)$ represent their respective histories up to epoch j [68].

A treatment strategy g is a policy that assigns treatment at each epoch. It may be defined as static if the same treatment is assigned to all individuals regardless of patient's history, or dynamic if treatment decisions depend on the patient's condition. For instance, propofol titration decisions based on current sedation depth are an example of a dynamic treatment strategy. Moreover, treatment strategies can be random if they introduce probability distributions over treatment values, or deterministic, if they refer to fixed assignments [68].

Given this formulation, the G-formula [66] computes the counterfactual mean outcome $E[Y^g]$ under a deterministic strategy g and with J time points as [67]:

$$E[Y^g] = \sum_{\bar{l}} E[Y \mid \bar{A} = \bar{a}^g, \bar{L} = \bar{l}] \prod_{j=0}^J f(l_j \mid \bar{a}_{j-1}^g, \bar{l}_{j-1}) \quad (2.4)$$

In essence, the formula determines the expected outcome under strategy g by averaging the outcome model over all possible covariate histories, weighted by the probability of each history occurring under the intervention. Notably, the G-formula fixes treatment at each epoch j to the value a_j^g , effectively replacing the observed treatment history with the counterfactual one, $\bar{a}^g$.

2.4.1.1 The Parametric G-formula

In the medical context, where data are continuous and high-dimensional, it is not feasible to sum over all covariate histories. Hence, the G-formula must be approximated. The parametric G-formula proposes to estimate counterfactuals by first fitting parametric regression models to each covariate [68] and then approximate intervention histories via Monte Carlo simulation.

The non-iterative conditional expectation (NICE) algorithm comprises three steps: parametric estimation, Monte Carlo simulation and calculation of mean under each intervention [68][69]. In the first stage, parametric regression models are fitted to the observed data for each component of the joint data distribution. Each time-varying covariate gets its own pooled user-specified model

that describes how they evolve over time as a function of other time-varying covariate and treatment history. A separate model is fitted to estimate the mean outcome given the full covariate and treatment history up to each epoch. The covariates included in the outcome model may comprise confounders identified from the causal graph through the adjustment set, as well as additional covariates included to improve the specification of the outcome model.

In the simulation phase, a new pseudo-population is generated, often larger than the original dataset. For each variable, at each epoch and for each simulated patient, simulated values are generated using the covariate models defined in the previous step. Baseline covariates are resampled from the observed data and left unchanged for all time-points. Additionally, for covariates subject to intervention, their values are assigned based on the user-defined decision rule [69].

The final step concludes the estimation process by computing the mean predicted outcome $E[Y^g]$ across all simulated patients in the new dataset created in the simulation phase [68][69].

2.4.2 Assumptions

A valid causal effect estimate resulting from the implementation of the G-formula to observational data rests on a set of assumptions [67].

The counterfactual consistency assumption dictates treatment unambiguity. In essence, if $\bar{A}_j = \bar{A}_j^g$ then $\bar{Y}_{j+1} = \bar{Y}_{j+1}^g$ and $\bar{L}_j = \bar{L}_j^g$ [68]. Given clear treatment definition, conditional exchangeability, also known as ignorability or no unmeasured confounding, determines that, at each epoch, the potential outcomes are independent of the observed treatment, conditional on the observed covariate and treatment history up to that point. Formally, for each epoch j , let Y_j^g and L_j^g respectively represent the counterfactual outcome and vector of confounders, for an individual assigned to a deterministic intervention $g(\forall j = 1, \dots, J)$ [70][68]:

$$(Y_{j+1}^g, \dots, Y_j^g) \perp (A_j, C_{j+1}) \mid \bar{L}_j = \bar{L}_j, \bar{A}_{j-1} = \bar{a}_{j-1}^g, C_j = Y_j = 0 \quad (2.5)$$

Finally, positivity establishes that, at each time point, every patient must have a non-zero probability of receiving treatment, that being any identified exposures, both static and dynamic, conditional on their observed covariate and treatment history [70][68]:

$$f_{\bar{L}_j, \bar{A}_{j-1}, C_j, Y_j}(\bar{L}_j, \bar{a}_{j-1}^g, 0, 0) > 0 \Rightarrow f_{A_j, C_{j+1} \mid \bar{L}_j, \bar{A}_{j-1}, C_j, Y_j}(a_j^g, 0 \mid \bar{L}_j, \bar{a}_{j-1}^g, 0, 0) > 0 \quad (2.6)$$

2.4.3 Causal Effect Estimand

The causal effects of a given treatment can be approximated at multiple levels. In this dissertation, the average treatment effect (ATE) estimand was utilized to quantify the estimated causal effects. ATE measures the population-level difference in outcomes under exposure to a treatment ($T=g$) compared to non-exposure ($T=0$). It may be adapted to estimate the difference in outcomes between a population in which everyone receives treatment ($T=g$) and the natural course of treatment ($T = \text{ref}$), representative of the observed outcome in the study population without an intervention

to alter exposure patterns [71]. It can be computed as the mean of individual treatment effects, such that

$$ATE = \frac{1}{N} \sum_{i=1}^N \tau_i \quad (2.7)$$

Where N is the total number of patients in the cohort and τ_i represents the individual treatment effect for patient i , defined as the difference between potential outcomes for patient i under treatment and reference conditions:

$$\tau_i = Y_i^g - Y_i^{ref} \quad (2.8)$$

2.5 Evaluation Metrics

Sensitivity and robustness analyses are commonly used to evaluate the reliability of causal effect estimates and assess their dependence on the causal assumptions [72].

2.5.1 E-values

E-values, defined as “the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates” already adjusted for [73], were used to quantify sensitivity to unmeasured confounding. In short, a large E-value indicates that only substantial unmeasured confounding could negate the observed effect estimate whereas an E-value close to 1 suggests a more fragile result, such that relatively weak unmeasured confounding would be sufficient to explain the estimated effect.

For an estimated risk ratio RR , the E-value is calculated as

$$E = RR^* + \sqrt{RR^*(RR^* - 1)} \quad (2.9)$$

where $RR^* = RR$ when $RR > 1$ and $RR^* = \frac{1}{RR}$ when $RR < 1$. Thus, the E-value is a measure of the minimum confounding strength, on the risk-ratio scale, required to move the observed estimate to the null.

2.5.2 Refutation Tests

To complement the causal analysis, refutation tests can be implemented to assess the stability of the estimated causal effects under data perturbations designed to challenge the underlying causal assumptions [72]. Such tests may include modifications to the treatment assignment, covariate structure or study population.

The placebo treatment refuter evaluates the dependence of an estimate on the observed treatment assignment by replacing it with a placebo treatment, for which no causal relationship with

the outcome should be observed [72]. Therefore, a refutation estimate close to the null supports the specificity of the observed treatment effect.

The random common cause refuter assesses the robustness of the estimated effects after the addition of an independent random variable as a confounder to the dataset [72]. Any substantial change to the estimate relative to the original estimated treatment effect, after its addition, suggests sensitivity to model specification.

Lastly, the data subset refuter investigates if the estimated effects are robust to changes in the study population by repeating the analysis on randomly selected subsets of the data. Hence, consistent estimates across subsets provide evidence that the observed effect is not driven by a particular percentage of the study population [72].

In short, the placebo refuter challenges the treatment-outcome relationship, the random common cause test focuses on model specification and the data subset refuter determines the generalizability of the estimate.

Chapter 3

Literature Review

Chapter 3 compiles prior research aligned with the themes explored in this dissertation. It reviews how Causal Inference methods have been applied to a multitude of critical care questions, with studies grouped by the primary methodological approaches identified in the literature: (1) propensity score-based methods, (2) G-methods, (3) reinforcement learning, (4) hybrid approaches and (5) federated learning. It then narrows its scope to evidence informing sedation practices in the intensive care unit, including divergent findings and remaining uncertainties. Moreover, it describes how Causal Inference methods have been used specifically for sedation and analgesia. Finally, the chapter considers the evidence supporting the methodological choices made in this dissertation and identifies the gap in current knowledge that this work aims to address.

3.1 Causal Inference in the ICU

The reach of Causal Inference has expanded from the social sciences to applications within the healthcare domain, much to the merit of the widespread adoption and availability of electronic health records [74] and a growing consensus that prediction alone is insufficient to guide clinical decision-making [48].

To date, a 2023 scoping review is the most recent and comprehensive overview of Causal Inference in healthcare, specifically in the intensive care unit. Smit and colleagues [75] selected 79 studies addressing time-varying interventions in the ICU and found that 61% of the included literature employed G-methods, with pronounced preference for inverse-probability-of-treatment weighting (IPTW), and reinforcement learning (RL) accounted for the remaining 39% of works. Moreover, studies were approximately equally divided between static (51%) and dynamic treatment regimes (49%) with a notable focus on three treatment categories: anti-inflammatory drugs, hospital-acquired complications and sedatives and analgesics. The review also emphasized significant inadequacies in the quality of reporting in the included articles, as only 38% of studies

reported all five target trial components (eligibility criteria, outcome, treatment strategies, follow-up period and analysis plan) and 9% acknowledged all three causal assumptions, which suggests that most published causal ICU studies cannot be validated.

Since Smit et al.'s review, propensity score-based methods continue to be a common choice for Causal Inference research in the healthcare space. Zhao et al. [76] used propensity score matching (PSM) to estimate the effect of dexmedetomidine exposure on mortality, length of ICU stay and successful weaning from MV, across ICU types and age groups, in mechanically ventilated sepsis patients. This work matched baseline covariates 1:1 using nearest-neighbour matching to reduce confounding, followed by fitting Cox proportional hazard models to assess the association between the exposure and the outcomes. They found that dexmedetomidine was associated with lower 28-day mortality, longer ICU stay and increased liberation from invasive mechanical ventilation (IMV), an effect present across age groups and ICU type. Yet, the exposure was defined as a binary indicator (dexmedetomidine use during ICU stay), which provides no insights into effects of dose variation or timing. Kim et al. [77] applied a similar propensity score approach to the same patient population, albeit from a different database (Korean Sepsis Alliance *versus* the previous MIMIC-IV cohort). These authors focused on evaluating the effect of the timing of mechanical ventilation on mortality, demonstrating that early (first day) MV, comparative to delayed MV, is associated with both lower ICU and in-hospital mortality, shorter ICU stays and decreased tracheostomy use. While PSM has its merits in time-fixed treatment scenarios, in the ICU, most treatment decisions are repeated throughout a patient's stay and conditioned on patient characteristics that are themselves affected by prior treatment decisions. This is known as treatment-confounder feedback and it prompts the use of more sophisticated methods like the G-methods, including IPTW, the parametric G-formula and G-estimation [66][75], to estimate causal effects using EHR.

Lee et al. [78] hypothesized that effect estimation, in this case, using IPTW, could be ameliorated by implementing deep sequence models, such as recurrent neural networks and transformer, to directly estimate propensity scores. Their work showed that these models led to more accurate estimates of treatment effects compared to the more standard feature processing approaches, highlighting its benefits in the context of time-varying confounding. However, both models were applied to synthetic and semi-synthetic datasets and, despite incorporating longitudinal data, evaluated the effect of a binary treatment at a single time point, which is a complexity mismatch with real clinical care. In a similar vein, Matos et al. [79] used machine learning-derived propensity scores, but replaced IPTW with a doubly robust estimator (Targeted Maximum Likelihood Estimation, TMLE) to investigate the impact of IMV and vasopressors on mortality in septic patients. TMLE showed no survival benefit of either invasive therapy but rather an association with increased mortality across septic patients, including those with cancer. Nevertheless, the doubly robust property of this estimator stands as a methodological improvement by protecting estimates against model misspecification of either the treatment or the outcome model.

Li et al. [80] and Yin et al. [81] proposed extending causal effect estimation into federated settings, to use decentralized data from multiple hospitals while preserving patient privacy. Li et al. presented FL-TTE, a federated learning-based target trial emulation framework in which

propensity scores are estimated locally and used for IPTW, followed by causal estimation via Cox proportional hazards models. Since only aggregated model parameters are communicated to a central server, there is no direct sharing of patient-level data. This method was validated using eICU Collaborative Research Database and MIMIC-IV cohorts and showed less biased estimates than conventional meta-analytic pooling. Yin et al. identified a related problem, noting that local model training may reintroduce confounding bias, even with accurate propensity estimates, if covariate-treatment relationships are not consistent across hospitals. Hence, they developed FED-IPTW, which forces de-correlation between covariates and treatment at the local and global level to reduce confounding in individual treatment effect estimation. When applied to estimating the causal effect of mechanical ventilation on survival in a sepsis cohort from the eICU Collaborative Research Database, the proposed approach outperformed existing baselines. Still, both methods inherit the limitations of the confounder set and the estimator they employ; thus, even if federation allows for increased data availability, it does not resolve potential confounding adjustment constraints.

In 2023, in what is a precursor of the methodological structure of this dissertation, Gani et al. [82] combined Causal Discovery with domain knowledge, provided by clinicians, to construct a structural causal model; identified a set of confounders via back-door adjustment and estimated the effect of oxygen therapy on ICU mortality using extreme gradient boosting. This framework extended prior work, which mostly relied on expert-defined causal-graphs; however, the authors acknowledge several important limitations such as treating oxygenation targets as static for the full ICU stay, which they attributed to the absence, at the time, of Causal Discovery methods to model time-varying treatment strategies.

Finally, Serpa Neto et al. directly addressed the concerns raised by Smit et al. [75] regarding study design by drawing on data from the TEAM trial [83] to emulate an hypothetical trial comparing mobilization “dose” treatment strategies (different combinations of duration and intensity) and using the parametric G-formula to estimate its effect. In this target trial emulation study, they simulated patient trajectory under discrete dose thresholds using time-varying G-formula, finding that prolonged mobilization was associated with increased 28-day mortality, in patients who only reached low mobility levels each day.

3.2 Sedation Management in the ICU

Despite sedation and analgesia being in the top three treatment categories identified in causal inference ICU research [75], the causal evidence for optimal sedation management remains insufficient.

The existing literature is dominated by randomised controlled trials (RCTs) where the suitability of different sedative drugs is compared, but their optimal use in practice is notoriously under-explored. The MENDS2 trial [39], comprised septic patients that were randomly administered either dexmedetomidine (0.2-1.5 mcg/kg/h) or propofol (5-50 mcg/kg/min) for up to 14

days, titrated to sedation depth targets, and found no significant difference between the two populations regarding outcomes such as delirium, coma, VFD-28 or 90-day mortality. Shehabi et al. [84] reached similar conclusions in the SPICE III trial, with a more general population of 4000 ICU patients, across 74 intensive care units. Furthermore, the SPICE III trial raised concerns over the use of dexmedetomidine without co-administration of another sedatives like propofol and demonstrated that this drug was associated with more haemodynamic adverse events [26]. However, dexmedetomidine has shown modest reduction in delirium incidence and time under MV for cardiac surgical patients, across 41 RCTs according to a systematic review by Heybati et al. [85]. RCT evidence generally agrees on a null result for drug choice regarding these sedatives, which supported PADIS guidelines encouraging the use of both propofol or dexmedetomidine as primary sedation agents [2].

If sedative choice has garnered considerable unanimity within the scientific community, defining an optimal sedation strategy is still an on-going debate, especially when it comes to sedation depth. In their meta-analysis, Stephens et al. [33] suggested that early light sedation was associated with lower mortality and duration reductions of mechanical ventilation, ICU stay and delirium incidence. A 2024 cohort study [34] challenged this finding and proposed a novel approach: an early deep-to-light sedation strategy, maintaining deep sedation targets for the first 48 hours and transitioning light sedation. The authors reported better outcomes when considering the dynamic strategy compared to the alternatives, suggesting that the timing and trajectory of sedation depth, and its personalization, may be more important than maintaining an average depth level. Additional studies have identified patients with severe conditions like ARDS and high ventilatory support needs where deeper sedation levels, paired with neuromuscular blockers, may be beneficial to stabilize their conditions and prevent further harm [86][87]. In contrast, the Olsen et al. [88] no-sedation trial, a sedative-free strategy showed slight improvements in cognitive outcomes when compared to light sedation, but no significant difference in mortality. Aitken et al. [30] summarized these inconsistencies best. When reviewing data from 26 studies, 8 RCTs and 18 cohort studies, they found that RCTs largely pointed toward a null benefit of maintaining lighter sedation levels, while observational studies reported improvements in some outcomes such as time to extubation, ICU length-of-stay and ventilator-related complications, yet no positive effects of a light sedation strategy in mortality or delirium. However, the authors acknowledge that part of this heterogeneity may be attributed to the lack of consistent definitions of "light" and "deep" sedation across the included studies. Still, ICU guidelines recommend light sedation targets, now clearly defined as RASS between -2 and +1, except in patient populations where its use is unadvised.

These conflicting results may be explained by variations in study design, clinical context, patient populations and clinical practices, [30] but they also highlight a methodological obstacle in retrospective sedation studies: the need to account for bias and confounding in observational data to guarantee valid results.

3.2.1 Causal Inference Methods to Study Sedation in the ICU

Sedation is a time-varying treatment, subject to treatment-confounder feedback, under which standard regression approaches produce biased estimates [75], motivating the use of Causal Inference methods. However, the application of such methods to sedation is scarce.

Table 3.1 comprises some key publications concerning the implementation of causal methods to answer clinical questions on sedation and analgesia.

Published research on sedation management has predominantly employed reinforcement learning (RL). Eghbali et al. [89] developed a deep RL approach, using MIMIC-IV data, to learn personalized dosing policies for the co-administration of propofol and fentanyl, extending on prior literature by modelling dual agents, accounting for dosage minimization trends in clinical practice, including haemodynamic variables in the reward function and approximating clinical settings by maintaining a continuous action space, that is, without discretizing medication dosing. A subsequent study by the same authors [90] continued to explore multi-agent RL strategies yet this time targeting delirium incidence in addition to sedation depth and haemodynamic targets. Similarly, Lee et al. [91] applied reinforcement learning to dexmedetomidine dosing and proposed an Artificial Intelligence (AI) model for Delirium prevention (AID) trained and externally validated across two cohorts, totalling 2805 ICU stays. They demonstrated that the estimated policy outperformed clinician decisions by +0.441 in the derivation cohort and +0.622 in the external validation cohort, but most interestingly, that the physician strategy had negative expected outcome across both cohorts. More recently, Romero-Hernández and Camara [92] applied offline RL to 47144 ICU stays, in an approach that accounted for partial observability of patient state by using recurrent neural networks to learn representations from histories of clinical observations and actions. Their study focused on modelling four continuous dosing actions based on eight sedative and analgesic agents, to learn policies that optimize drug prescription for pain reduction and survival benefit. They found that policies that strictly reward pain relief without considering long-term outcomes (i.e. mortality) were associated with increased mortality in the ICU, which contrasts with previous studies that did not penalize mortality in the reward function [93] [89].

Additional observational studies in this domain mostly follow propensity-score-based methodologies. Muriel et al. [94] applied marginal structural models with inverse probability of treatment weighting to estimate the causal effects of sedatives and analgesics during non-invasive positive pressure ventilation on its failure. A decade later, Wu et al. [95] used PSM in a cohort of critically ill surgical patients to evaluate the association between early deep sedation and one-year post-hospital mortality, resulting in a positive association between these variables after matching. The authors also found that sedative agent choice, specifically the use of midazolam or propofol rather than dexmedetomidine, age, gender, body mass index, concurrent disease and illness severity are all risk factors for patient survival. However, the analysis did not account for the dynamic treatment-confounder relationship during the patient's ICU stay.

Table 3.1: Causal approaches to estimate the effects of different sedatives and analgesics in the ICU. Studies marked * are present in the scoping review by Smit et al. (2023) [75]; remaining studies consist of additional relevant work identified in this dissertation.

Reference	Method	Treatment / outcome	Dataset	Study design	Key finding
Sinzinger et al. (2005)* [96]	RL	Propofol/sedation depth measured by BIS	Simulated data	Dynamic	RL agent successfully controlled patient consciousness, displaying improved consistency and adaptability to changing targets compared to conventional controllers
Muriel et al. (2015)* [94]	IPTW	Sedatives and analgesics / need for MV	842 patients across 322 ICUs	Static	The individual use of either sedative or analgesics had no effect on the outcome but their combined used was associated with increased need for invasive MV
Padmanabhan et al. (2015)* [97]	RL	Propofol / maintenance of BIS and MAP targets	Simulated data	Dynamic	The RL agent maintained the clinical target values, for both variables, for the majority of the time and patients
Padmanabhan et al. (2017)* [98]	RL	Propofol and remifentanyl / sedation depth measured by BIS	Simulated data	Dynamic	The RL-based controller optimized the concurrent dosing of sedatives and analgesics with synergistic effects
Lopez-Martinez et al. (2019)* [93]	RL	Morphine / Maintenance of clinical targets and pain scores	6843 patients from MIMIC III	Dynamic	The RL agent provided personalized morphine dosing recommendations to minimize pain scores and maintain target vital sign values
Eghbali et al. (2021)* [89]	RL	Propofol and fentanyl / maintenance of SAS and MAP clinical targets	1757 patients from MIMIC-IV	Dynamic	RL policy outperformed clinicians on both SAS and MAP targets
Ribba et al. (2022)* [99]	RL	Propofol / BIS target	Simulated data	Dynamic	Coupling RL with PK-PD modelling may increase dosing accuracy
Eghbali et al. (2023) [90]	RL	Propofol, midazolam and fentanyl / maintenance of RASS and MAP clinical targets and delirium	1,346 patients from MIMIC-IV	Dynamic	RL policy outperformed clinicians on every target
Lee et al. (2024) [91]	RL	Dexmedetomidine / delirium	2416 patients from Supreme 2.0 and 270 patients from SICdb	Dynamic	RL-optimized dexmedetomidine dosing reduced predicted delirium
Wu et al. (2025) [95]	PSM	Early sedation (RASS) / mortality	7135 patients from Taichung Veterans General Hospital database	Static	Early deep sedation is associated with increased one-year mortality in critically ill patients
Cui et al. (2026) [100]	Causal Forest Double ML	Propofol and fentanyl / postoperative ICU admission	67134 surgical procedures from the Irvine Medical Center	Static	Identified high-risk and optimal combinations of sedatives and analgesics
Romero-Hernández et al. (2026) [92]	RL	Sedatives and analgesics / pain and mortality	30087 patients from MIMIC-IV	Dynamic	Pain-only policies are associated with increased mortality, while joint policies (pain and mortality) are associated with lower mortality

Abbreviations: BIS, bispectral index; IPTW, inverse probability of treatment weighing; MAP, mean arterial pressure; ML, machine learning; PK-PD, pharmacokinetics and pharmacodynamics; RASS, Richmond Agitation–Sedation Scale; RL, reinforcement learning; SAS, Riker Sedation-Agitation Scale; SICdb, Salzburg Intensive Care database.

In 2026, Cui et al. [100] applied a counterfactual causal machine learning framework to concomitant propofol and fentanyl dosing using perioperative EHR, discovering high-risk dose combinations that increase postoperative ICU admission (fentanyl dose above 5 mcg/kg while propofol dose remains below <1 mg/kg), identifying optimal dosing intervals equal to fentanyl doses between 3.5 to 4.0 mcg/kg and propofol doses within 1.25 to 4.25 mg/kg and demonstrating the potential of causal dose-response analysis for sedation and analgesia.

3.3 Dissertation Positioning

Despite the clear predominance of reinforcement learning applied to dynamic treatment regimes in sedation and analgesia, these approaches have important limitations. Common critiques to the use of RL in Causal Inference include data inefficiency, heterogeneity in how clinical problems are formulated as Markov decision processes and a lack of standardized evaluation metrics and baselines for comparison across studies [101][102]. Furthermore, the interpretability of treatment recommendations derived from RL can be a considerable challenge, especially when "black box" models are employed [102]. In clinical settings, however, interpretability and explainability have been shown to increase clinician's confidence in AI-based decision-support [103].

In order to address some of these limitations, recent works have explored the integration of causality concepts (i.e. causal graphs) into reinforcement learning frameworks, originating causal reinforcement learning [102]. Nevertheless, the main goal of any RL study is to learn and optimize treatment policies, which is an inherently distinct task from estimating the causal effects of well-defined treatment strategies in the presence of time-varying confounding.

Given the current paradigm of sedation in the ICU, where evidence on the effects of sedative agent choice on mortality remains limited and sedation depth targets are still contested, the predominance of RL and propensity-score-based approaches exposes a clear gap for interpretable and benchmarked methods to evaluate the effects of sedative dosing strategies on clinical outcomes.

This dissertation aims to address that gap directly by employing the parametric G-formula, a validated method suited for longitudinal causal effect estimation under time-varying confounding [66] [67] and which, to the author's knowledge, has not been applied to sedation management, to study the causal effects of propofol dosing strategies in critically ill mechanically ventilated patients.

Chapter 4

Methods

Chapter 4 details the workflow employed in this dissertation, covering four distinct steps: Data Preparation, Causal Discovery, Causal Inference and Robustness and Sensitivity Analysis. It first expands on patient selection and the construction of baseline and longitudinal datasets to capture patient initial status and trajectory, respectively. Subsequently, the Chapter elaborates on how this data was processed before presenting the ensemble-based Causal Discovery approach used to retrieve the causal structure of the treatment-outcome relationship and identify an adjustment set. It then describes the Causal Inference analysis, where, using the parametric G-formula, the causal effects of five sedation strategies, both static and dynamic, were estimated. Finally, it explains how results were assessed via E-value computation and refutation tests.

4.1 Data

4.1.1 Data source

The clinical data were sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, version 3.1. [104] which includes the de-identified medical records of patients admitted to both the emergency and intensive care unit (ICU) departments of the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, between the years of 2008 and 2022. Upon completion of the Collaborative Institutional Training Initiative Program’s “Data or Specimens Only Research” training (certification report no. 73526447), the PhysioNet database was accessed via Google BigQuery under a data use agreement. This database received a waiver of informed consent from the International Review Board at the BIDMC due to the implemented protection measures, which enabled its publicly accessible status.

4.1.2 Cohort Definition

The cohort definition strategy is illustrated in the diagram in Figure 4.1.

Initially, a total of 94,458 ICU admissions were retrieved from the MIMIC-IV database. Patient eligibility was then refined to patients that met the following criteria:

- Patients under mechanical ventilation (first episode);
- Age ≥ 18 years old;
- ICU length of stay ≥ 24 hours;
- Admission to an eligible ICU type: Medical Intensive Care Unit (MICU), Surgical Intensive Care Unit (SICU), Medical/Surgical Intensive Care Unit (MICU/SICU), Trauma Surgical ICU (TSICU) or Coronary Care Unit (CCU);
- At least one recorded propofol infusion during the ventilation window;
- Sepsis meeting the Sepsis-3 clinical definition.

Within this subset of patients, individuals were excluded if they satisfied the conditions below:

- Repeated ICU admissions;
- Use of extracorporeal membrane oxygenation (ECMO);
- Admission to the Cardiovascular Intensive Care Unit (CVICU) or non-ICU units (e.g. Post-Anesthesia Care Unit and other step-down units);
- Tracheostomy at or within 24 hours of mechanical ventilation start;
- Glasgow Coma Score (GCS) = 3.

The eligible ICU types were identified by taking into consideration the clinical indications for propofol administration and infusion duration. For instance, CVICU was excluded because propofol use in this unit is often associated with short-term postoperative sedation, following cardiac surgery, rather than prolonged ICU sedation management [105]. Furthermore, early analysis of the study cohort indicated that patients admitted to CVICU were ventilated, on average, for 42.8 hours, compared to the MICU, where mechanical ventilation had a mean duration of 145.6 hours. The timing of propofol initiation while under MV also differed between units, with a mean interval of 1.5 hours from intubation to propofol initiation in the CVICU and 13.5 hours in the MICU.

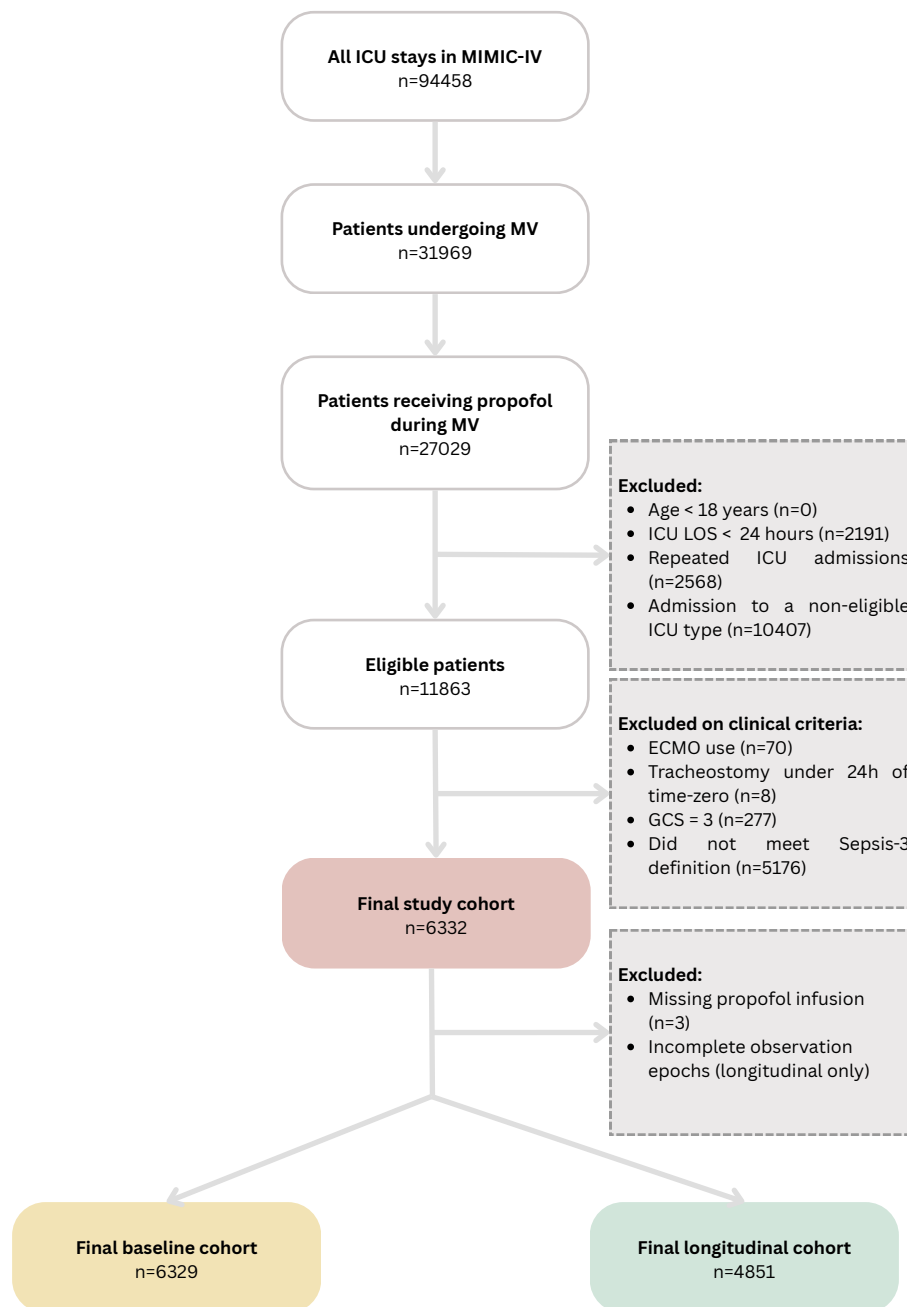


Figure 4.1: Study cohort flowchart. Time-zero definition: time stamp of each patient’s first propofol infusion while under mechanical ventilation. Abbreviations: ECMO, extracorporeal membrane oxygenation; GCS, Glasgow Coma Score; ICU, intensive care unit; LOS, length of stay; MV, mechanical ventilation.

Regarding clinical exclusion criteria, patients with pre-existing tracheostomy at the start of the observation period were excluded to limit the cohort to those receiving mechanical ventilation via an endotracheal tube. Moreover, patients receiving ECMO were not included in the analysis as this form of respiratory support differs from standard MV and may alter drug pharmacokinetics through circuit sequestration [106]. Finally, patients with a Glasgow coma score equal to 3 were

excluded, since this degree of neurological impairment may be indicative of acute brain injury and may, therefore, require specialised sedation management [89]. Besides, neurocritical patients may also hinder the evaluation of sedation depth using validated scales like the Richmond Agitation-Sedation Scale (RASS).

After applying all criteria, the study cohort comprised 6332 ICU stays. Nonetheless, further exclusions were applied to comply with methodological requirements of the analysis. In this context, three patients with missing propofol infusion rate data at the start of the observation period were excluded from the study cohort, resulting in the final baseline cohort. Moreover, the longitudinal study required patients to have complete follow-up, that is to be ventilated throughout the observation window (24 hours); hence, patients who died or were extubated prior to the end of the analysis were not accounted for. As such, the final longitudinal cohort comprised 4851 patients.

4.1.3 Datasets

The analytic datasets created from the MIMIC-IV v3.1. database refer to the patients encompassed in the study's cohorts and are indexed to time-zero (t_0), defined as the individual's first propofol infusion while under mechanical ventilation (MV). The treatment of interest was, therefore, the propofol infusion rate. The measured outcome was ventilator-free days at 28 days (VFD-28), defined as the number of days alive and free from invasive MV during the 28 days following t_0 . Patients who died within this time frame or remained under MV beyond the 28-day time period were assigned VFD-28 equal to zero. The number of VFD-28 of the remaining cohort was computed as the difference between the total amount of days (28) and the number of days spent on MV from t_0 until extubation.

4.1.3.1 Baseline Dataset

The cohort's baseline status was captured in a dataset comprising 55 variables extracted once per patient from the data available during the 24-hour window before t_0 . These covariates were informed by clinical literature on mechanical ventilation and sepsis within a critical care setting [89] [107] and cover demographic features such as age, weight and gender, illness severity indicators like Sequential Organ Failure Assessment (SOFA) score and Acute Physiology Score III (APS III), comorbidities, prior medications, patient vitals, laboratory values and haemodynamic condition, as well as respiratory function measurements.

When multiple measurements of a given variable existed in the specified time window, the aggregation strategy varied based on the clinical role of the covariate and is summarized in Table 4.1. Most continuously monitored variables, such as laboratorial results or vital signs, were extracted as the mean of all recorded values. Exceptions were made for lactate and creatinine, which were extracted as the peak value within the time interval. These variables serve as illness severity proxies, since elevated lactate levels may reflect tissue hypoperfusion and metabolic

stress, while increased creatinine values may indicate impaired renal function or acute kidney injury [108] [109]. Additionally, urine output was represented as the cumulative volume produced during this time window to capture renal function beyond what is indicated by creatinine levels. Summing urine output over the 24 hours prior to t_0 is consistent with clinical criteria for assessing acute kidney injury, in which urine production is evaluated relative to body weight and duration over defined time periods ranging from 6 to 24 hours [110].

Ventilator parameters, such as the fraction of inspired oxygen (FiO_2), positive end-expiratory pressure (PEEP), and plateau pressure, as well as sparse clinical assessments like GCS, were obtained by considering the most recently documented value. The former because ventilator settings reflect physician-controlled parameters rather than direct physiological measurements, and the latter since the assessment preceding a sedation decision was considered the most informative measure of patient neurological status. Similarly, the closest values to time-zero of blood gas variables, including pH, the partial pressure of arterial oxygen (PaO_2) and the partial pressure of arterial carbon dioxide (PaCO_2), were deemed the most representative of patient physiological state at the start of propofol sedation management.

All binary exposure variables were coded as 1 if there was a valid record within the considered time window, defaulting to 0 if no record was present and treated as non-exposure. Unlike midazolam exposure, dexmedetomidine and fentanyl mean rates were included as continuous features due to literature supporting propofol and fentanyl co-administration [89] and the substitution of propofol as a sedative agent for MV patients for dexmedetomidine supported by the 2025 SCCM Focused Update [36].

Table 4.1: Aggregation map for baseline covariate extraction.

Aggregation method	Variables
Mean	Fentanyl infusion rate, dexmedetomidine infusion rate, norepinephrine-equivalent dose, MAP, heart rate, SpO_2 , respiratory rate, mean airways pressure, hemoglobin, WBC, platelets, ALT, AST, bilirubin, albumin, anion gap, glucose, bicarbonate, sodium, potassium, BUN, calcium, RASS
Maximum value	Creatinine, lactate
Last value (<i>closest to t_0</i>)	GCS, FiO_2 , PEEP, plateau pressure, ventilator mode, pH, PaO_2 , PaCO_2
Sum	Urine output
Occurrence/ Exposure	Vasopressor use, renal replacement therapy, midazolam, neuromuscular blockers use, SBT, SAT, liver disease

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; FiO_2 , fraction of inspired oxygen; GCS, Glasgow Coma Score; MAP, mean arterial pressure; PaCO_2 , partial pressure of arterial carbon dioxide; PaO_2 , partial pressure of arterial oxygen; PEEP, positive end-expiratory pressure; RASS, Richmond Agitation-Sedation Scale; SAT, spontaneous awakening trial; SBT, spontaneous breathing trial; SpO_2 , peripheral oxygen saturation; WBC, white blood cell count.

4.1.3.2 Longitudinal Dataset

Strictly conditioning on baseline covariates would be insufficient to address the time-varying confounding problem of clinical decision-making [67] [70]. Therefore, the longitudinal nature of sedation practices in the ICU was captured in a dataset of time-varying covariates, following patient trajectory for 24 hours, corresponding to six 4-hour epochs to match nursing practices for sedation depth measurements [37].

The extraction of the treatment variable, as the average propofol rate within each 4-hour epoch, was complemented by two additional features. The first ran the cumulative sum of propofol infusions to inform posterior dosing decisions by considering accumulation effects and the risk of PRIS, and the second corresponded to a binary flag to model epochs where propofol rate was equal to 0 mcg/kg/min. Furthermore, a binary censoring covariate was included to determine if a patient remained eligible for follow-up as ascertained by the cohort's inclusion criteria. In essence, this variable indicated if the patient was no longer ventilated, either due to extubation or death, within the 24 hours after t_0 , effectively censoring patient trajectory.

The choice of relevant time-varying covariates to extract aimed to cover most clinical domains and align with the identified adjustment set spanning neurological status, sedation depth, administration of concurrent medications, namely sedatives and analgesics, haemodynamic variables (e.g. mean arterial pressure and heart rate) and the use of vasopressors (captured by the continuous norepinephrine equivalent dose) due to propofol's known cardiovascular effects. It also encompassed respiratory variables such as ventilator settings (FiO_2 , PEEP, ventilation mode) and weaning attempts. Notably, SATs were excluded on account of a definitional overlap with sedation targets rather than genuine weaning attempts. Yet, with SBT being a frequent pre-cursor to SAT, the covariate capturing SBT signals was deemed sufficient. End of follow-up outcomes, including ventilator-free-days at 28 days and mortality at 28 days were carried forward per patient.

Similar to the aggregation strategy present for the baseline dataset, any variable that was continuously monitored and, therefore, charted more frequently, included an averaged measurement of the values documented within each epoch, and any variable with sparser measurements or corresponding to physician set parameters was reported as the last observation within a given epoch.

4.1.3.3 Merged Dataset

Finally, the parametric G-formula allows for the inclusion of both baseline and time-varying covariates provided in one singular input. As such, a merged dataset was created by combining the baseline covariates comprising the adjustment set with the longitudinal dataset, which included the longitudinal adjustment set covariates as well as additional variables to better represent the patient's clinical state. The final dataset contained one row per patient per epoch, with each row comprising the time-varying measurements at that epoch and the fixed baseline covariate profile for that patient.

4.1.4 Data Processing

Data processing was conducted separately for the baseline and longitudinal covariates although both datasets were first cleaned to remove variables with zero variance, more than 80% missing data or with high risk of collinearity with other covariates. Furthermore, in both approaches, the variables were assigned a role as identifiers, outcomes, treatment variables, binary, categorical or continuous covariates, to facilitate imputation.

4.1.4.1 Baseline Data

As previously established, the baseline dataset comprised data obtained in the 24 hours before time zero. While identifiers, binary variables and certain covariates such as SOFA score or Charlson Comorbidity Index (CCI) presented no missing values, either due to the availability of records or the coalescing logic during extraction, categorical variables, namely ICU type and ventilation mode, were imputed using the most frequently observed category prior to ordinal encoding. As an example, ventilation mode was grouped into three categories including volume-controlled (0), pressure-controlled (1) and spontaneous/pressure-support ventilation (2).

Continuous covariates, on the other hand, were standardized using z-score normalization to prevent variables with larger numerical ranges from dominating the Euclidean distance computations, before being imputed using k-nearest-neighbour (KNN) imputation with $k=5$. Following imputation, the continuous variables were transformed back to their original scale and clipped to physiological ranges.

The baseline cohort characteristics were then compared across treatment groups, categorized by propofol infusion rate, as described later in this chapter. Standardized mean differences (SMDs) were calculated for a subset of baseline covariates, with the lowest-dose group used as the reference. An absolute SMD above 0.1 was deemed representative of possible baseline imbalance [111].

4.1.4.2 Longitudinal Data

The longitudinal dataset included time-varying covariate measurements per epoch and per patient during the follow-up period.

Data missingness was addressed by first implementing forward-fill within each patient's trajectory to carry forward the last observed value, when such was documented. This aimed to reflect clinical practice where a lack of charted measurement does not necessarily dictate an absence of treatment but most frequently an absence of treatment change. Subsequently, the remaining missing values were imputed using KNN with $k=5$, similar to the baseline data.

Ventilation mode, a categorical variable, was imputed using per-patient mode when multiple observations were recorded within an epoch, falling back to the cohort-level mode in cases where a patient had no valid charted values.

4.2 Causal Discovery

4.2.1 Algorithm Selection

In total, three Causal Discovery algorithms were selected to recover the causal structure of the propofol rate and VFD-28 relationship, using the observational data extracted from MIMIC-IV. PC, fGES and BOSS were implemented using Carnegie Mellon University’s Py-Tetrad package (version v0.1.1), the Python interface to the Tetrad Causal Discovery software [112].

The combination of constraint-based and score-based methods aimed to leverage their individual strengths and increase result robustness. The Peter-Clark algorithm provided flexible conditional independence testing while fGES and BOSS provided global score-based searches that were less sensitive to error propagation [113]. Hence, parallelizing these approaches not only reduced computation time but also allowed for complementary edge detection.

The PC-stable algorithm variant applied the degenerate Gaussian conditional independence test with a significance threshold of $\alpha=0.01$ while fGES and BOSS used the degenerate Gaussian score with a penalty discount of 1. Both the degenerate Gaussian score and test accommodate mixed data types [62], making it suitable for ICU datasets comprising binary variables such as attempts at SAT and continuous covariates like propofol rate. Moreover, each algorithm was run with 500 bootstrap resamples of the dataset, with replacement and resample size equal to 100% of the original data, using Tetrad’s built-in resampling. Edge frequencies were logged across the 500 bootstrap runs per algorithm, resulting in a stability criterion that reflects how reliably each edge was recovered.

Due to the computational demands of this task, the Causal Discovery analysis was run on the high-performance computing (HPC) cluster provided by Fraunhofer Portugal AICOS, with 16 CPUs and 64 GB of memory, using the baseline covariate dataset ($n=6,329$ patients and 48 variables).

The chosen algorithms support the specification of additional inputs, such as metadata and background knowledge encoding variable ordering constraints. The following subsection provides further detail on the latter.

4.2.2 Background Knowledge

In this study, domain knowledge was incorporated as temporal tier system that guided the causal structure learning process by constraining the search space and preventing physiologically implausible edge directions. To do so, baseline variables were assigned to six tiers that reflect both their temporal and causal order (Figure 4.2). In specific, covariates were characterized as demographic, if inherent to the patient, admission data when measured at hospital intake or early ICU stay and ICU data when recovered during their stay in this department. Moreover, composite scores like SOFA and APS III were placed in the fourth tier as they are derived from measurements obtained previously, such as the values included in the preceding tier. Finally, the last two tiers reflect the starting propofol infusion rate and the ventilator-free days at 28 days for each patient.

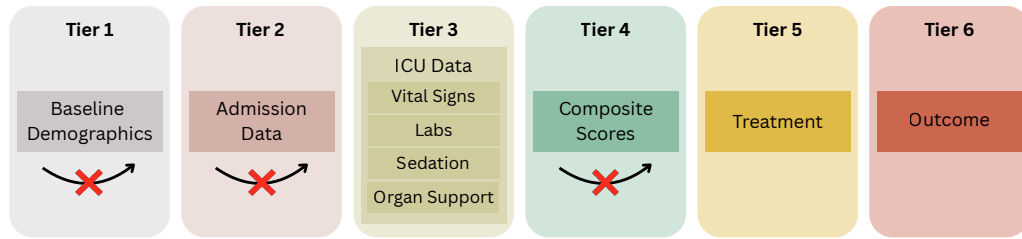


Figure 4.2: Temporal tiers and forbidden edges identified in the study dataset.

At this stage, edges that violated the defined temporal order or encoded relationships known to be non-causal (e.g. connections between demographic variables) were prohibited to prevent spurious associations and backwards causation. Furthermore, edges grounded in prior clinical knowledge, namely the connection between RASS and propofol infusion rate and propofol rate and the outcome were specified and enforced in the model.

The basis for their inclusion is extensively reported in the literature and clinical practice. For instance, sedation guidelines in critical care promote sedative titration to achieve a target sedation depth measured by validated scales such as the Richmond Agitation-Sedation Scale. Additionally, not only is the existence of a causal relationship between the treatment and outcome paramount in a valid causal graph, but also randomised controlled trials support the causal pathway between propofol dosing and mechanical ventilation duration and ventilator-free days [39].

4.2.3 Ensemble Approach

An ensemble of the three Causal Discovery algorithms increased confidence in the recovered causal structure by evaluating edge consistency across a total of 1500 bootstraps. This strategy aggregated the edge frequencies of 500 bootstraps per algorithm, such that, for each candidate edge, a combined frequency score, with a maximum value of 3, was calculated. Edges whose combined frequency exceeded the threshold of 1.5, that is, that were present in the majority of bootstraps and, thus, had a higher than chance probability of reflecting a causal relationship, were retained in the consensus DAG.

The ensemble strategy introduced directed cycles, which violates the assumption of acyclicity in a DAG, prompting their removal by eliminating the least-frequent edge. The final acyclic ensemble graph was saved as a .dot file for visualisation and identification of the adjustment set.

4.2.4 Adjustment Set Identification

The adjustment set Z was derived from the ensemble DAG by identifying the parents of propofol infusion rate and excluding any mediators of the treatment-outcome relationship, complying with Pearl's back-door criterion [59] described in Chapter 2.

Conditioning on the variables in Z blocks all back-door paths between propofol rate and ventilator-free days, removing confounding and allowing for the estimation of the causal effects of propofol dosing strategies on VFD-28 rather than capturing spurious associations. Therefore,

the identified adjustment set informed the baseline covariates to be included in the parametric G-formula models.

4.3 Causal Inference

The parametric G-formula was implemented using the `pygformula` package developed by CAUSALab [114], which applies the NICE estimator to estimate counterfactual outcomes under hypothetical treatment strategies and compute the causal effects of each intervention from observational data.

4.3.1 Covariate Specification

Covariates were separated by their respective functions in the model. In essence, baseline covariates (“basecovs”) were passed as a fixed set and kept unchanged throughout all epochs and time-varying covariates were handled individually. First, each time-varying covariate was assigned a type according to their distribution, such as binary, normal, zero-inflated normal or categorical and then, each covariate was modelled based on this pre-specified type, such that, for instance, binary variables were modelled using logistic regression and continuous (assumed normal distribution) covariates were modelled using linear regression.

The “covmodels” argument specified the predictors included in these models. It consisted of a user-defined string for each covariate, including the relevant variables and their history functions used to predict the covariate values at the next epoch. In this context, the `pygformula` package incorporates three pre-coded history functions: `lag`, `cumavg` and `lag_cumavg`, encoding the lagged, cumulative average and lagged cumulative average value of a covariate at a given time point (usually the previous epoch), respectively.

Furthermore, knowing that model specification is order-dependent, the covariate models followed the causal structure identified in the previous step, such that variable dependency followed the direction encoded in the DAG, avoiding circular dependencies within the same epoch and with lagged values capturing the temporal dependencies.

Overall, “covmodels” were described as diverging combinations of lagged values of the covariate to simulate, lagged and concurrent values of additional time-varying covariates, baseline variables and linear or squared time variables (“epoch_num” and “square_epoch_num” respectively).

4.3.2 Outcome Model

The outcome model predicted ventilator-free days at 28 days as a function of the full treatment and covariate history. To avoid misspecification, the user-defined model did not include all time-varying covariates but rather a select subset of variables believed to be the most predictive of VFD-28 according to the literature and the causal graph. The influence of baseline covariates and time were introduced via “basecovs” and the linear time variable, similar to the covariate models and as described above.

4.3.3 Model Assessment

Model misspecification was assessed by evaluating the agreement between the parametric natural course estimate, which is the Monte Carlo mean estimate of VFD-28 under the treatment patterns observed in the dataset, and the non-parametric natural course estimate, computed directly from the observed data. Furthermore, the user-defined covariate models were iteratively refined based on the natural course plots supported by the `pygformula` package, to visually reduce discrepancies between the parametric and non-parametric estimates at each epoch and improve model fit. Minimal deviations, in the decimal order, were considered adequate model specification.

4.3.4 Treatment Strategies

This study focused on five distinct interventions: four static, where treatment is fixed at every epoch in spite of the patient’s clinical evolution, and one dynamic. The static treatment strategies were defined based on the literature, highlighting doses that increase risk for propofol-related infusion syndrome, as follows:

Table 4.2: Static propofol infusion rate strategies.

Strategy	Propofol infusion rate
No propofol	0 mcg/kg/min
Maintenance dose	$0 < \text{dose} \leq 50$ mcg/kg/min
Above-maintenance dose	$50 < \text{dose} \leq 67$ mcg/kg/min
High dose	> 67 mcg/kg/min

The zero-dose category was included to capture epochs during which propofol was not administered, either due to a sedative agent change or on-going weaning protocols.

The dynamic intervention’s design followed current ICU sedation guidelines, which recommend propofol titration to light sedation targets, that can be assessed with tools such as the Richmond Agitation-Sedation Scale (RASS), reflecting scores within -2 and +1. Therefore, under this strategy, the propofol dose category assigned to each 4-hour epoch was adjusted according to the mean RASS value from the preceeding epoch. For the first epoch, the dose category adjustment was based on the baseline RASS value at t_0 . The dose category remained unchanged if RASS was within the target sedation range, was reduced by one category if RASS was below -2, potentially indicating oversedation, or was, otherwise, increased by one category if RASS surpassed +1, suggesting agitation or inadequate sedation.

The natural course estimate captured the outcomes under the observed treatment decisions in the dataset, without imposing an intervention, serving as a reference for the performance of the alternative treatment strategies.

4.4 Robustness and Sensitivity Analysis

CAUSALab's pygformula package provides built-in non-parametric bootstrapping for retrieving confidence intervals (CI) for all G-formula estimates. For every 500 bootstrap replicates, patients were resampled with replacement and the complete G-formula pipeline was re-run, resulting in bootstrap distributions of estimated treatment effects for each comparison with the reference. 95% CIs were constructed using the percentile method, taking the 2.5th and 97.5th percentiles of this distribution [115]. Bootstrapping and associated simulation were parallelised across 8 CPU cores to reduce computation time.

Beyond the aforementioned confidence intervals, result robustness was assessed by evaluating sensitivity to unmeasured confounding, to data changes and violations of the causal assumptions.

E-values were computed from the estimated mean ratios (MR), considering that the rate ratio may be regarded as analogous to the risk ratio for non-negative continuous outcomes [73]. In such case, the MR can be directly substituted in the E-value formula as

$$E = MR^* + \sqrt{MR^*(MR^* - 1)} \quad (4.1)$$

where $MR^* = MR$ when $MR > 1$ and $MR^* = \frac{1}{MR}$ when $MR < 1$. Sensitivity to unmeasured confounding was equally approximated by calculating E-values for the 95% confidence interval bound closest to the null value.

The placebo treatment refuter was implemented by extracting each patients treatment trajectory, randomly permuting these trajectories and assigning them to different patients. This approach preserves the temporal structure of treatment decisions while breaking their correspondence with each patient's covariate and outcome history. The treatment effect was then re-estimated using the permuted trajectories.

To overcome pygformula's implementation restrictions for time-varying covariates, the random common cause test introduced an additional randomly generated baseline variable, with values independently drawn from a standard normal distribution for each patient. This new random covariate was then added to "basecovs" and included in the outcome model during parametric G-formula estimation.

Finally, the data subset refuter randomly selected 80% of the study populations and re-estimated the treatment effects using each subset to assess result stability.

Each refutation test was run for 10 iterations and their results were summarized by the mean estimated treatment effects across repetitions. The evaluation criterion depended on the refutation test performed, such that, for the placebo treatment refuter, an estimate was considered to pass if the estimated effect collapsed to zero, while for both the random common cause and data subset refuters, robustness was assessed based on whether the refuted estimate remained within the 95% confidence interval of the original estimate.

Chapter 5

Results

Chapter 5 compiles the findings of every stage of the causal pipeline. It explores the study’s cohort characteristics and clinical profiles, the results of the ensemble approach combining three distinct Causal Discovery algorithms, presenting the adjustment set for the propofol dose-VFD28 relationship. Moreover, causal effect estimates for each treatment strategy defined in Chapter 4, alongside their bootstrap confidence intervals, are reported and compared. Robustness of these estimates to violations of the methodological assumptions is evaluated. The Chapter concludes by discussing the findings through both a methodological and clinical lens.

5.1 Cohort Characteristics

A total of 6329 patients, meeting the eligibility criteria, were retrieved from the MIMIC-IV database. The primary analysis followed 4851 patients that remained under mechanical ventilation (MV) for a total of 24 hours from time-zero (defined as the first propofol exposure while under MV), using six 4-hour epochs. A summary of the cohort’s baseline characteristics is provided in Table 5.1.

The cohort was predominantly male (58.5%) with a mean age of 61.19 ± 16.99 years and an average weight of 85.77 ± 26.22 kg. As is consistent with a subset of critically ill patients, severity indicators were mildly elevated. The primary cohort presented moderate burden of concurrent disease on 1-year mortality, given by a Charlson comorbidity index of 4.59 ± 3.01 [116], and mean SOFA and APS III scores of 7.07 ± 4.07 and 56.29 ± 24.14 , respectively, reinforcing moderate-to-high mortality risk and illness severity. The mortality rate for this population was 23.6% and the average amount of days free from ventilation at 28 days was 16.84 ± 10.75 . Vasopressor use was seen in 35.9% of patients at baseline, with a relatively low average norepinephrine-equivalent dose of 0.07 ± 0.17 mcg/kg/min [117], and 20.6% of the cohort had diagnosed mild or severe liver disease, potentially affecting drug clearance pathways.

Table 5.1: Summary of baseline characteristics and outcomes of the study cohort.

Variable	Overall	Missing, %
Population		
N patients	4,851	—
Demographics		
Age, years	61.19 ± 16.99	0.0
Male sex, %	58.5	0.0
Weight, kg	85.77 ± 26.22	0.3
Severity and Comorbidities		
APS III score	56.29 ± 24.14	0.0
SOFA score	7.07 ± 4.07	0.0
Charlson Comorbidity Index	4.59 ± 3.01	0.0
GCS	14.39 ± 1.85	26.0
Liver disease, %	20.6	0.0
Medication Exposure		
Midazolam prior, %	20.4	0.0
RASS	-2.13 ± 1.77	39.2
SAT prior, %	52.2	0.0
Vasopressor use, %	35.9	0.0
Norepinephrine-equivalent dose, mcg/kg/min	0.07 ± 0.17	64.1
Haemodynamics and Respiratory		
MAP, mmHg	82.91 ± 14.88	7.2
Heart rate, bpm	91.84 ± 18.92	6.0
FiO ₂ , %	65.58 ± 24.17	8.4
PEEP, cmH ₂ O	6.76 ± 3.00	9.4
Treatment		
Hours from ventilation to propofol initiation	10.31 ± 37.93	0.0
Propofol rate at t_0 , mcg/kg/min	34.29 ± 20.38	0.0
ICU Type, %		
MICU	32.9	—
MICU/SICU	14.3	—
SICU	22.3	—
TSICU	22.7	—
CCU	7.7	—
Outcomes		
VFD-28, days	16.84 ± 10.75	0.0
28-day survival, %	76.4	0.0

Overall values are present as mean $\pm$ standard deviation except where noted. The absence of a recorded observation of binary variables was registered as a lack of exposure. Abbreviations: APS III score, Acute Physiology Score III; SOFA score, Sequential Organ Failure Assessment score; GCS, Glasgow Coma Scale; RASS, Richmond Agitation-Sedation Scale; SAT, spontaneous awakening trial; MAP, mean arterial pressure; FiO₂, fraction of inspired oxygen; PEEP, positive end-expiratory pressure; MICU, medical intensive care unit; SICU, surgical intensive care unit; TSICU, trauma/surgical intensive care unit; CCU, Cardiac care unit; VFD-28, ventilator-free days at 28 days.

Regarding haemodynamic conditions, patients had an average MAP equal to 82.91 ± 14.88

mmHg, falling within the target range for adequate perfusion (65-85 mmHg) [89], and a mean heart rate of 91.84 ± 18.92 bpm. Additionally, respiratory parameters remained within normal range, indicative of stable ventilated patients. However, among mechanically ventilated patients with sepsis, ventilator settings are reassessed at multiple time points and adjusted in accordance with oxygenation targets and the patient's lung condition, often starting with a more protective approach [118].

Prior sedative exposure was common for this subset of patients, with 20.4% of patients having received midazolam. 52.2% of the cohort experienced a documented spontaneous awakening trial in the 24 hours prior to time zero, which aligns with clinical practice as introduced by Chapter 1: "sedation vacations", SBTs and SATs are encouraged by current sedation guidelines. Furthermore, sedation depth, as measured by the Richmond-Agitation Sedation Scale, averaged a score of -2.13 ± 1.77 at baseline, representative of a lightly to moderately sedated patient cohort.

With respect to treatment exposure, the longitudinal trajectory of propofol dosing evolved across epochs. At time zero, 75.3% of patients were administered propofol doses up to 50 mcg/kg/min, 17.9% within the range of 50-67 mcg/kg/min and 6.7% received higher doses associated with deeper sedation and increased risk of PRIS (>67 mcg/kg/min) [44]. Over the primary analysis time period, the frequency of each treatment strategy was 25.1% for patients following a no propofol intervention, 61.4% for patients receiving standard maintenance doses, 10.4% for propofol rates between 50 mcg/kg/min and 67 mcg/kg/min and 3.1% for the higher-dose strategy. The elevated percentage of patients whose most frequent intervention was no propofol may indicate that they have been switched to an alternative sedation agent, such as dexmedetomidine, or it may reflect clinical weaning practices and preparation for awakening trials. Among epochs with active propofol infusion, the mean propofol rate was 34.29 ± 20.38 mcg/kg/min.

In sedation observational studies, confounding may arise from inherent differences between patients who receive different sedative doses. There is a common confounding example implicating the correlation between sicker patients and the administration of higher sedative doses and worse outcomes. Interestingly, this study's cohort displays the opposite behaviour, presenting a clinical picture where more severely ill patients are under a lower propofol dose protocol, as evidenced by Table 5.2.

Table 5.2: Standardized differences across propofol dose categories

Variable	Maintenance (0-50 mcg/kg/min)	Above maintenance (50-67 mcg/kg/min)	SMD	High (>67 mcg/kg/min)	SMD
Demographics					
Age, years	63.27 ± 16.49	55.87 ± 16.87	−0.444*	52.14 ± 16.77	−0.669*
Male sex, %	56.6	63.6	0.142*	66.6	0.206*
Weight, kg	84.33 ± 26.07	89.79 ± 25.41	0.212*	91.21 ± 28.24	0.253*
Severity and Comorbidities					
APSO score	58.47 ± 24.18	50.12 ± 21.75	−0.363*	48.26 ± 25.14	−0.414*
SOFA score	7.42 ± 4.11	6.22 ± 3.68	−0.307*	5.46 ± 3.89	−0.490*
Charlson Comorbidity Index	4.90 ± 2.99	3.76 ± 2.90	−0.387*	3.30 ± 2.73	−0.558*
Medication Exposure					
RASS	−2.12 ± 1.75	−2.11 ± 1.81	−0.004	−2.27 ± 1.77	−0.087
Haemodynamics and Respiratory					
MAP, mmHg	81.84 ± 14.37	86.00 ± 16.29	0.271*	86.72 ± 14.93	0.334*
Treatment timing					
Hours from ventilation to propofol	11.81 ± 39.36	6.86 ± 37.04	−0.129*	2.67 ± 16.358	−0.303*
Outcome					
VFD-28, days	16.06 ± 11.05	19.14 ± 9.45	0.300*	19.43 ± 9.31	0.330*
28-day survival, %	73.3	85.4	0.302*	87.4	0.361*
N patients	3,655	870	—	326	—

Overall values are reported as mean ± standard deviation unless stated otherwise. The standardized mean differences (SMDs) were computed relative to the maintenance-dose intervention (0-50 mcg/kg/min), as such, the first SMD column compares the above-maintenance group (50-67 mcg/kg/min) with the maintenance group and the second compares the high-dose group (>67 mcg/kg/min) with the reference. Following Austin (2009) [111] conventions, absolute SMD values > 0.1 indicate the possibility of significant baseline covariate imbalance and are marked with *. Abbreviations: APSIII, Acute Physiology Score III; SOFA, Sequential Organ Failure Assessment; RASS, Richmond Agitation-Sedation Scale; MAP, mean arterial pressure; VFD-28, ventilator-free days at 28 days.

5.2 Causal Discovery Results

The final ensemble DAG identified 249 edges among the 48 nodes retained from the PC, BOSS and fGES graph skeletons at the specified threshold. Additionally, the required edges linking RASS and propofol rate and the latter with VFD-28 were preserved in the final graph, in accordance with the literature.

The adjustment set derived from the DAG (Figure 5.1) identified eleven baseline variables which block all backdoor paths between propofol rate and ventilator-free days at 28 days. Categorically, these can be grouped into demographics and admission data, spanning age, sex, weight and ICU type, respiratory variables, namely the fraction of inspired oxygen (FiO_2), sedation-related confounders like sedation depth (given by RASS), prior SAT attempts and prior exposure to midazolam, and finally, severity indicators, such as acute physiology score III (APSIII score), mean arterial pressure (MAP) and timing from the commencement of MV until the first infusion of propofol.

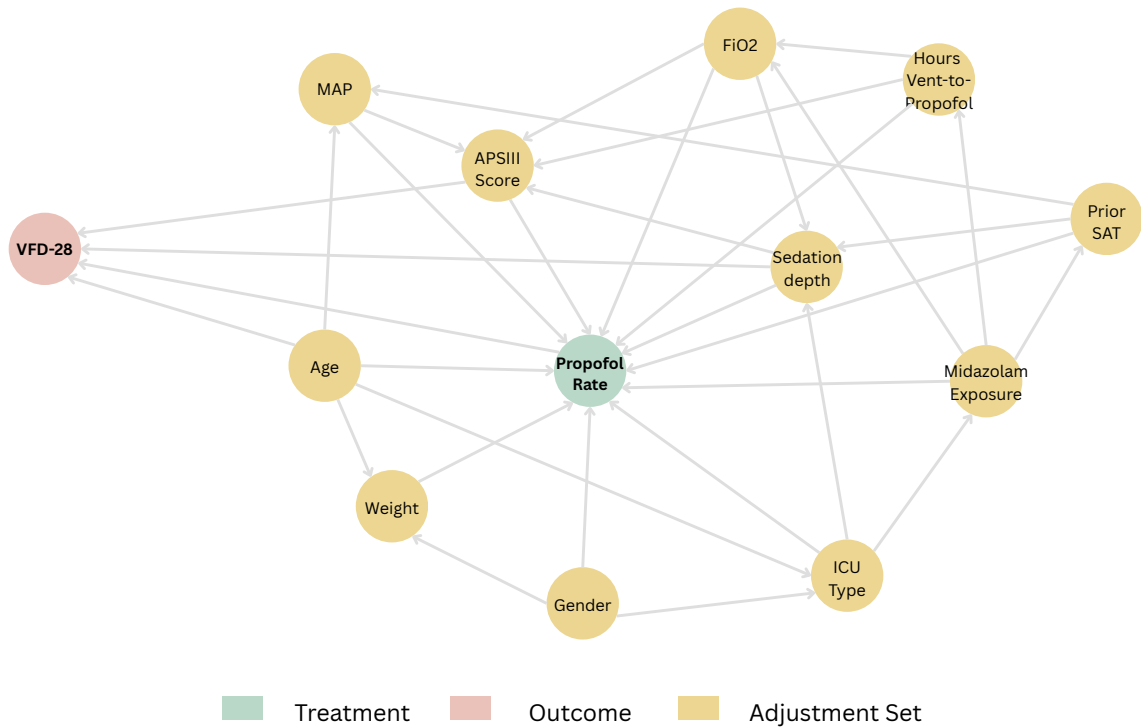


Figure 5.1: Subgraph of the consensus causal DAG illustrating the adjustment set for the relationship between propofol rate and VFD-28. The treatment is highlighted in green, the outcome in red and the confounders in yellow. Abbreviations: APSIII score, acute physiology score III; FiO_2 , fraction of inspired oxygen; MAP, mean arterial pressure; SAT, spontaneous awakening trial; VFD-28, ventilator-free days at day 28.

5.3 Causal Inference Results

Table 5.3 compiles the results of the mean counterfactual VFD-28 estimates under the pre-specified propofol dosing strategies. Moreover, Figure 5.2 depicts the individual covariate models, representing the correspondence between the simulated trajectories and the non-parametric means at the epoch-level, which were obtained via pygformula's natural course diagnostic plots.

The parametric G-formula mean estimate under the natural course (NC) yielded 16.43 VFD-28 (95% CI 15.92-16.74), corresponding to a 0.41 days gap when compared to the non-parametric estimate of 16.84 days. The no propofol strategy resulted in the lowest estimated VFD-28 at 16.00 days (95% CI 15.46-16.43), suggesting that following this policy may result in substantial harm relative to observed clinical practice (-0.43 days (95% CI -0.67, -0.18)). In contrast, the continuous infusion of propofol doses between 0 mcg/kg/min and 50 mcg/kg/min, the most frequent dosing range in the present cohort, obtained an estimated ventilator-free days at 28 days of 16.62 days (95% CI 16.02-17.01), the closest value to the NC. The estimated effect of this intervention implies an increase of 0.18 (95 % CI -0.04, 0.40) extubated days for patients following this approach, yet the 95 % confidence intervals include zero, which suggests some uncertainty regarding this effect. Notably, propofol doses slightly above maintenance values produced the best outcomes, corresponding to 17.07 days (95% CI 16.27-17.64) and an estimated effect of an additional 0.64 VFD-28 days (95% CI 0.09, 1.17) compared to NC. Higher doses (>67 mcg/kg/min) resulted in a decrease of 0.12 ventilator-free days (95% CI -1.23, 0.80). The two static treatment strategies related to higher propofol rates, 50-67 mcg/kg/min and >67 mcg/kg/min, present wider confidence intervals, consistent with the lower number of samples in these dose ranges (n=870 and n=326, respectively). Furthermore, the dynamic intervention, following clinical guidelines to titrate based on sedation depth to achieve light sedation targets, produced an estimated VFD-28 of 16.17 days (95% CI 15.66-16.53), with a mean difference of -0.26 (95% CI -0.39, -0.11), falling below natural course estimates.

Table 5.3: Estimated ventilator-free days at 28 days (VFD-28) under natural course and hypothetical propofol dosing interventions using the parametric G-formula (500 bootstrap replicates).

Intervention	NP mean	G-formula mean (95% CI)	Mean Difference	MD 95% CI
Natural course (reference)	16.84	16.43 (15.92, 16.74)	—	—
Always no propofol†	—	16.00 (15.46, 16.43)	-0.43	(-0.67, -0.18)
Maintenance (0–50)	—	16.62 (16.02, 17.01)	+0.18	(-0.04, 0.40)
Above maintenance (50–67)†	—	17.07 (16.27, 17.64)	+0.64	(0.09, 1.17)
High-dose (>67)	—	16.31 (15.06, 17.20)	-0.12	(-1.23, 0.80)
RASS-guided†	—	16.17 (15.66, 16.53)	-0.26	(-0.39, -0.11)

MD = mean difference in VFD-28 days relative to natural course (reference). Positive values indicate more ventilator-free days. † indicates 95% confidence interval (CI) excludes zero. Dose ranges in mcg/kg/min.

Addressing model fit, a visual assessment of individual covariate models, demonstrated proximity between the trajectory of observed data and the parametric estimate at each epoch, for most

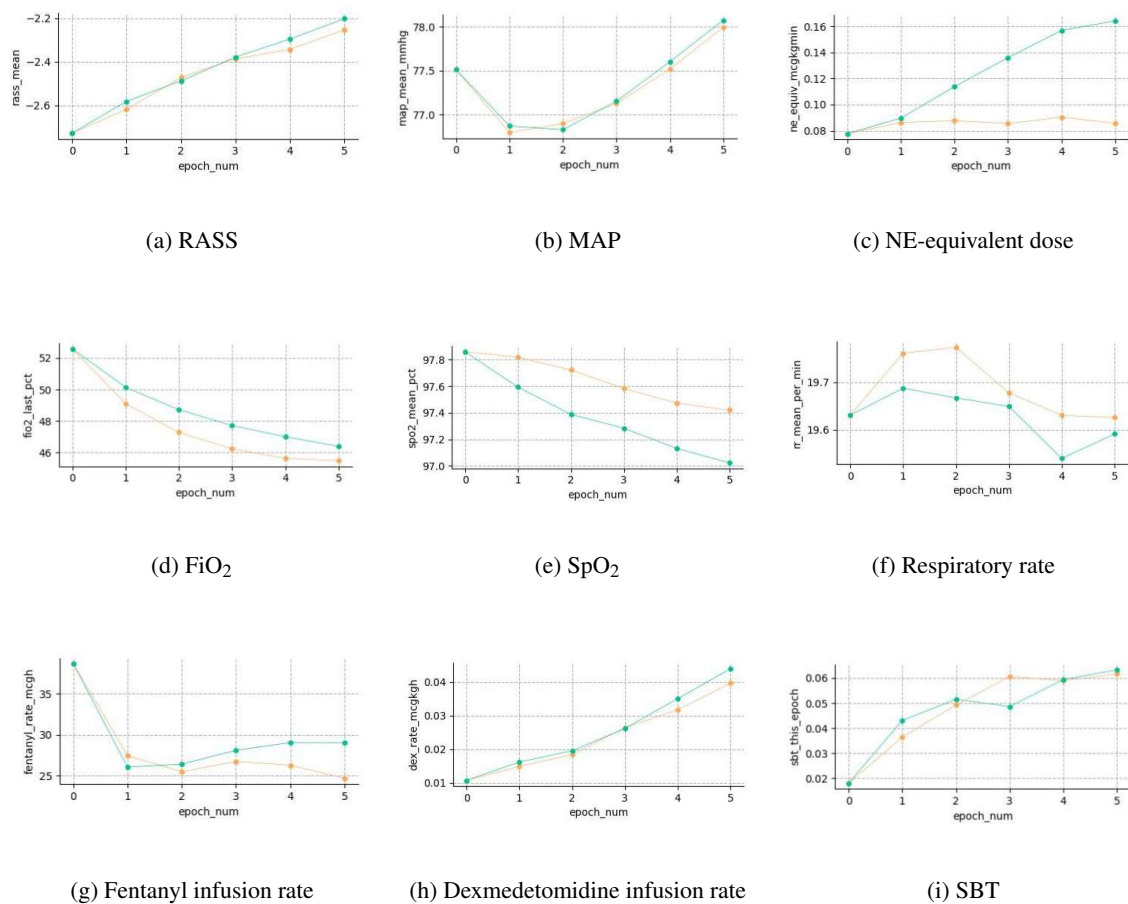


Figure 5.2: Comparison of simulated covariate trajectories (green) and non-parametric means (orange) across six epochs, for nine key time-varying covariates. Plots (a-i) encompass sedation depth measured by the Richmond Agitation-Sedation Scale, haemodynamics (mean arterial pressure and norepinephrine-equivalent dose), respiratory support (fraction of inspired oxygen, peripheral oxygen saturation, respiratory rate), and sedative/analgesic dosing (fentanyl, dexmedetomidine), as well as spontaneous breathing trials occurrence.

variables.

Sedation-related covariates, such as RASS, dexmedetodimine rate and the administration of opioids (fentanyl) displayed adequate fit compared to the non-parametric trends, with parametric mean overestimating the causal effects at later epochs.

The hemodynamic variable of mean arterial pressure (MAP) closely followed the observed trajectory across all six epochs. On the other hand, although the equivalent norepinephrine dose showed strong correspondence with the non-parametric means in the first two epochs, beyond this point it diverged significantly, progressing upwards despite the reference remaining mostly stable.

Finally, all respiratory covariates captured the overall trend but demonstrated minor deviations from the non-parametric patterns. For instance, the simulated trajectory of the fraction of inspired oxygen (FiO_2) was consistently overestimated across the time-window. Conversely, both the peripheral capillary oxygen saturation (SpO_2), and the respiratory rate were underestimated throughout the six epochs. The parametric estimates for attempted spontaneous breathing trials, the precursors to SATs, aligned with the non-parametric estimates with the exception of the fourth epoch. Nevertheless, the differences between the simulated and observed means for these covariates were small, with variations at the decimal level.

Despite multiple iterations to re-specify the covariate models, by adapting covariate history functions and adding other predictors, fit did not improve beyond what is presented in this chapter.

5.4 E-values and Refutation Tests

The estimates robustness to residual confounding was captured by E-values included in Table 5.4. All interventions reported small point E-values ranging from 1.09 to 1.24 and corresponding confidence interval bound results from 1.03 to 1.26. These findings imply that an unmeasured confounder associated with both propofol dosing and VFD-28 by a risk ratio of 1.08, at the lower limit, would be enough to reduce the lower confidence bound to null. Therefore, the interpretation of the causal effect of propofol dosing is subject to analysis under possible weak unmeasured confounding.

Table 5.4: Sensitivity analysis including E-values and refutation test results by intervention.

Intervention	MD (95% CI)	E-value (point / CI bound)	Placebo Treatment	Random Common Cause	Data Subset
Always no propofol	−0.43 (−0.67, −0.18) [†]	1.19/1.08	0.002	−0.43	−0.36
Maintenance (0–50)	+0.18 (−0.04, +0.40)	1.12/1.00	−0.024	0.18	0.17
Above maintenance (50–67)	+0.64 (+0.09, +1.17) [†]	1.24/1.10	0.051	0.64	0.52
High (>67)	−0.12 (−1.23, +0.80)	1.09/1.00	0.203 [‡]	−0.12	−0.07
RASS-guided	−0.26 (−0.39, −0.11) [†]	1.14/1.07	0.007	−0.26	−0.22

MD = mean difference in VFD-28 days *versus* natural course, included for context with full estimates depicted in Table 5.3. [†] 95% CI excludes zero. [‡] Placebo mean falls outside original 95% CI. Point E-values are computed at the point estimate, while CI bound E-values are calculated at the confidence interval bound closest to the null.

Refutation test results demonstrated that the estimated causal effects of all interventions were stable to the introduction of random common causes and to subsampling. The estimate associated with the continuous infusion of high propofol doses (>67 mcg/kg/min) did not collapse to zero once treatment assignment was randomized, contrary to the remaining treatment strategies. This finding, alongside the wide 95% CI for this intervention, suggest that the identified causal effect may reflect a random variation rather than a stable result.

5.5 Discussion

At its core, this dissertation aimed to investigate how different propofol dosing strategies influence ventilator-free days at 28 days among ICU patients with sepsis, and whether these effects can be estimated from electronic health record data while adjusting for confounding inherent to critical care. However, the findings provided only partial support for this premise. The Causal Discovery approach identified a clinically relevant adjustment set and subsequent Causal Inference analyses estimated heterogeneous causal effects across treatment strategies. The absence of propofol and higher propofol doses were associated with fewer ventilator-free days, whereas maintenance and slightly above-maintenance dosing strategies showed more favourable outcomes. Nevertheless, the robustness analyses and methodological constraints limited the extent to which these results could support definitive conclusions.

Causal Discovery identified a back-door adjustment set comprising eleven clinically consistent potential confounders of the treatment-outcome relationship. These variables included individual characteristics such as age, sex and weight which may affect sedative titration. Propofol is dosed on a weight-adjusted basis and older patients are more susceptible to its hypnotic effects, often requiring lower infusion rates to reach sedation targets [119]. Moreover, the intensive care unit type to which patients are admitted to may reflect diverging patient populations and clinical practices [120]. Confounding introduced by illness severity was captured by the acute physiology score

III (APS III score), mean arterial pressure (MAP) and fraction of inspired oxygen (FiO₂), which collectively indicate physiological derangement, existing haemodynamic instability and respiratory support needs prior to treatment initiation. Furthermore, hours from ventilation to propofol start may reflect the rate of patient deterioration that prompts sedation. Lastly, midazolam exposure, spontaneous awakening trials (SAT) attempts and mean Richmond Agitation-Sedation Scale (RASS) describe sedation trajectory before time-zero and may directly impact the initial propofol infusion rate.

The back-door adjustment set was derived from an ensemble of three algorithms – PC, BOSS and fGES –, enhancing the reliability of its results comparatively to the use of a single algorithm or selecting confounders based on theory alone. It is, nonetheless, important to acknowledge that the present approach only included baseline data, which cannot fully characterise the dynamic relationships between covariates, especially when it concerns to time-varying treatments subject to treatment-confounder feedback, as feedback loops violate acyclicity in static analysis and are, therefore, eliminated. Furthermore, the Causal Discovery findings must be validated by clinical experts by reviewing the final graph and adjustment set to guarantee its clinical plausibility and that no important relationships are omitted.

The simulated estimates under all treatment regimes yielded an equal split of positive and negative effects, with the strongest evidence of potential benefit being associated with doses above maintenance rates (50-67 mcg/kg/min), corresponding to an increase of 0.64 (95% CI 0.09, 1.17) ventilator-free days, which remained stable across placebo treatment, random common cause and data subset perturbations. This intervention sits in between standard clinical practice and higher propofol doses, with the latter being a known risk factor for propofol-related infusion syndrome (PRIS) when administered consistently for prolonged periods of time. Its apparent benefit could be explained by altered drug pharmacokinetics in patients with sepsis, as physiological dysregulation may alter volume of distribution and hepatic function, which could lead to higher dose requirements to reach comparable sedation depth targets to those in non-septic ICU patients [26] [24]. This theory appears to be supported by the mean RASS scores reported in Table 5.2, which indicate that the above-maintenance group maintains light sedation ranges, according to PADIS 2018 guidelines [2], despite receiving higher infusion rates. It is important to acknowledge that the highest performing treatment strategy resulted in a moderate absolute benefit of approximately 15 additional ventilator-free hours. Nevertheless, even small increments could have considerable implications in patient recovery and survival, as well as in ICU resource utilization and in the reduction of the daily cost of mechanical ventilation, which, as previously established (Chapter 1) drives global ICU burden.

Although the estimated causal effect of high-dose propofol (>67 mcg/kg/min) suggested potential harm comparatively to observed clinical practice, the estimate was imprecise, with a 95% CI spanning two VFD-28 and crossing zero. This uncertainty may, however, be partly justified by the relatively small sample size (n=326 at baseline) of this treatment group. Additionally, this was the least stable estimate across refutation tests and was the only one for which the placebo treatment result did not collapse to zero. Thus, this estimate was considered unreliable. The lower

estimated effect relative to the previous intervention is compatible with the aforementioned concerns regarding its association with PRIS risk [44]; nevertheless, the 24-hour observation window of this study does not allow one to directly infer this risk from its findings. As such, future research should follow patient trajectories over longer time windows to evaluate the safety of continuous high-dose propofol exposure.

Propofol dosing strategies following maintenance rates produced a near-null result (+0.18 days 95% CI -0.04, 0.40), with slight crossing over zero, potentially indicating close alignment with real sedation practices and meeting expectations for the most common observed treatment regime. Conversely, the no propofol strategy revealed a stable estimated negative effect (-0.43 days 95% CI -0.67, -0.18), which, despite existing research into non-sedation approaches [88], is supported by sedation guidelines endorsing propofol as one of the standard sedative agents for mechanically ventilated patients [2].

Lastly, the RASS-guided strategy was associated with an estimated reduction of 0.26 VFD-28 (95% CI -0.39, -0.11), which appears to contradict current guidelines. This discrepancy can be, in part, attributed to the fact that the MIMIC-IV database includes ICU admissions from 2008 to 2022, a period during which sedation practice evolved substantially. Across these years, new evidence supported the gradual implementation of sedation interruptions, SBTs and SATs, and early deep sedation became associated with worse outcomes. Hence, comparing the estimated effect of a RASS-guided treatment strategy, which reflects a relatively novel approach to sedation management, with the observed natural course throughout the total study period, may be misleading. Another possible explanation is the dynamic regime itself, as implemented in this dissertation. The simple model adjusts propofol dose category based on the RASS value at each 4-hour epoch and does not allow for titration within treatment category, as may occur in routine clinical care, especially within maintenance dose ranges. Therefore, the estimated effect does not necessarily indicate that a RASS-based propofol titration strategy is harmful, but may instead reflect methodological limitations. These could potentially be ameliorated by modifying the dynamic treatment rule to include additional clinical context that informs RASS target decisions, such as MAP, on account of propofol's haemodynamic effects [89][97], and by restricting ICU admission to data entries representative of current clinical practices.

These findings are, however, subject to residual confounding as captured by the E-value sensitivity analysis. The low E-values across all intervention estimates, ranging from 1.09 to 1.24, indicate that a relatively weak unmeasured confounder could explain away the observed effects [73]. Thus, despite the study having been designed to estimate causal effects through adjusting for the back-door criterion confounders, the obtained estimates, cannot be interpreted as true causal effects, given the likelihood of unmeasured confounding.

This is particularly relevant in the context of potential covariate imbalance and the observed severity gradient across treatment groups documented in Table 5.2. Patients receiving higher propofol infusion rates had consistently lower illness severity and comorbidity burden, as indicated by the sequential organ failure assessment (SOFA) score, APS III and Charlson comorbidity index (CCI): when compared to the maintenance-dose group, the absolute SMDs for SOFA were

0.307/0.490, for APS III 0.363/0.414 and for CCI 0.387/0.558 for the above-maintenance and high-dose groups, respectively. Given these imbalances, the possibility of residual confounding by indication cannot be excluded.

Beyond the limitations already addressed, this work is affected by additional methodological constraints.

The primary analysis included 4851 patients out of the total 6329 patients in the cohort because the parametric G-formula software (pygformula) required complete longitudinal records, effectively excluding patients who died or were extubated during the observation window and retaining patients who were sufficiently stable to remain under mechanical ventilation throughout follow-up. Although this exclusion affects both patients with increased illness severity and those with better clinical status, it may introduce selection bias. Hence, the estimated effects may have limited generalisability beyond this patient subset.

Furthermore, the parametric G-formula has a recognized limitation: the reliance on well specified, user-defined models for the time-varying covariates and outcome. Despite most covariate models producing simulated trajectories that closely approximated the non-parametric means, others, such as norepinephrine-equivalent dose model, could not adequately capture the trends in the observed data. This may be, in part, attributed to the current implementation, which only included linear covariate histories and linear or quadratic time functions, restricting its ability to model more complex dose-response relationships.

The parametric G-formula also relies on Monte Carlo simulation to approximate the treatment effects, thus, the NICE estimator is subject to Monte Carlo simulation error, which decreases as number of simulated individuals increases [121]. The present analysis included 500 bootstraps using a simulated sample size equal to the total cohort size, so residual error can be present. Additionally, errors in the simulated covariate history at a given epoch may propagate to subsequent time points because simulated histories are used to predict future covariate values, and, therefore, enhance the deviations in the simulated trajectories across the observation window. These limitations are inherent to the parametric G-formula and cannot be overcome by model re-specification, even with further developments such as introducing censoring, competing events or restrictions, some of which are only currently available for survival outcomes. Hence, a potential avenue for future work would be to re-estimate the causal effects using a doubly robust estimator such as targeted maximum likelihood estimation to better protect estimates against model misspecification.

Despite exploratory work to derive a policy tree, the study's cohort did not display sufficient heterogeneity to enable the identification of optimal and clinically actionable treatment policies. Therefore, further research could extend the causal pipeline by including a more diverse patient population or focusing on patient subgroups with greater treatment effect differences. Additionally, alternative or even composite outcomes could be considered, such as maximizing VFD-28 while minimizing ICU length of stay, although this approach would require G-formula estimates for each outcome included in the policy.

Finally, the MIMIC-IV database includes admissions to a single medical unit in the United

States, the Beth Israel Deaconess Medical Center; therefore, these findings may not be generalizable to different hospitals with diverging patient populations or sedation practices.

Chapter 6

Conclusion

As electronic health record data becomes more available and complex, critical care research must go beyond statistical associations and use it to recover personalized insights into patient care. To that end, understanding the causal relationships between clinical interventions and patient outcomes is paramount to translating data into actionable knowledge that can inform clinical decision-making.

This dissertation aimed to develop a Causal Inference pipeline combining Causal Discovery and parametric G-formula estimation to recover the causal effects of alternative propofol dosing strategies on VFD-28 in mechanically ventilated patients with sepsis, using MIMIC-IV patient records.

Overall, this work demonstrated that above-maintenance propofol dosing (50-67 mcg/kg/min) was the best performing intervention out of the five treatment regimes tested, yielding +0.64 days (95% CI 0.09, 1.17) relative to observed clinical practice. It was also the only positive effect estimate whose confidence interval excluded zero and that remained stable across all refutation tests. The maintenance dosing strategy (0-50 mcg/kg/min) aligned the natural course of treatment, and a sedation protocol without propofol was associated with worse ventilatory outcomes by delaying extubation by 0.43 days (95% CI -0.67, -0.18). Moreover, high propofol doses appeared to be associated with worse outcomes, as initially expected and consistent with propofol-related infusion syndrome risk, although the 24-hour study time window and the imprecise causal estimate for this intervention prevented drawing definitive conclusions. Interestingly, the dynamic RASS-guided strategy, which most closely reflects 2018 guidelines for the Prevention and Management of Pain, Agitation/Sedation, Delirium, Immobility, and Sleep Disruption in Adult Patients in the ICU, by adapting sedative titration to a sedation depth target, produced a negative VFD-28 effect relative to standard clinical care. However, methodological limitations, such as the use of a simple model that prohibits propofol titration within a treatment category could explain this apparent harm. Nevertheless, causal interpretation of these results is conditional on their low E-values and on possible remaining confounding by indication.

Taken together, these findings indicate that administering propofol rates within the range of 50 mcg/kg/min to 67 mcg/kg/min could be a beneficial strategy for this patient population; yet,

further work is required to support this claim before changes to clinical practice are considered. Moreover, by identifying a clinically consistent back-door adjustment set through Causal Discovery and obtaining a robust estimate for the above-maintenance finding, this work provides a basis for subsequent investigations into optimal propofol dosing strategies in mechanically ventilated sepsis patients.

6.1 Future Work

This work fulfilled its stated objectives; however, future research could extend its scope by addressing some of the limitations of the present approach.

For instance, one could expand the Causal Discovery analysis by employing longitudinal methods, such as PCMCI [122] or LPCMCI [123], to account for time ordering beyond background knowledge, model lagged effects and treatment-response feedback and better characterize the dynamic relationships between clinical variables, without breaking the causal assumptions (e.g. acyclicity). Given the E-values observed in this study and the prevalence of latent confounding in medical data, LPCMCI-based approaches would be particularly relevant to accommodate unmeasured confounders. Additionally, since the identified causal relationships and adjustment set were not independently validated, future clinical validation could provide further support to the findings.

Another avenue for improvement could be targeting methodological limitations caused by the current parametric G-formula implementation like integrating censoring and competing events for continuous outcomes (which is not yet supported by the `pygformula` package) in order to overcome the complete-epoch requirement and allow the full cohort to be considered in the estimation process. Furthermore, one could explore the inclusion of custom machine learning models and custom history functions to improve data fit. However, this will not fully resolve model misspecification concerns as the variables, lagged covariates and cumulative exposures to be included in the model would still be user-defined. As such, an alternative approach, using a doubly robust estimator like targeted maximum likelihood estimation, could be a protective strategy against treatment or outcome model misspecification.

Regarding interventions, increasing the complexity of the dynamic treatment strategies to represent the interactions between various time-varying covariates, like propofol dose, sedation depth and mean arterial pressure, for example, would better reflect how treatment decisions are made in clinical practice.

Finally, given that the lack of patient heterogeneity in the observed cohort prevented the identification of optimal treatment policies, subsequent investigations into this topic should consider broader patient populations, clinically relevant patient subgroups and composite outcomes.

Ultimately, any future developments should aim to better reflect the complexity of critical care so that the estimated causal effects can more accurately inform actionable sedation policies.

References

- [1] Basil Varkey. Principles of clinical ethics and their application to practice. *Medical Principles and Practice*, 30(1):17–28, June 2020.
- [2] John W. Devlin, Yoanna Skrobik, Céline Gélinas, Dale M. Needham, Arjen J. C. Slooter, Pratik P. Pandharipande, Paula L. Watson, Gerald L. Weinhouse, Mark E. Nunnally, Bram Rochwerg, Michele C. Balas, Mark van den Boogaard, Karen J. Bosma, Nathaniel E. Brummel, Gerald Chanques, Linda Denehy, Xavier Drouot, Gilles L. Fraser, Jocelyn E. Harris, Aaron M. Joffe, Michelle E. Kho, John P. Kress, Julie A. Lanphere, Sharon McKinley, Karin J. Neufeld, Margaret A. Pisani, Jean-Francois Payen, Brenda T. Pun, Kathleen A. Puntillo, Richard R. Riker, Bryce R. H. Robinson, Yahya Shehabi, Paul M. Szumita, Chris Winkelman, John E. Centofanti, Carrie Price, Sina Nikayin, Cheryl J. Misak, Pamela D. Flood, Ken Kiedrowski, and Waleed Alhazzani. Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the icu. *Critical Care Medicine*, 46(9):e825–e873, September 2018.
- [3] Hannah Wunsch, Jason Wagner, Maximilian Herlim, David H. Chong, Andrew A. Kramer, and Scott D. Halpern. Icu occupancy and mechanical ventilator use in the united states. *Critical care medicine*, 41(12):2712–2719, December 2013.
- [4] Temesgen Ayenew, Mihretie Gedfew, Mamaru Getie Fetene, Belayneh Shetie Workneh, Animut Takele Telayneh, Afework Edmealem, Bekele Getenet Tiruneh, Guadie Tewabe Yinges, Addisu Getie, and Mengistu Abebe Meselu. Prevalence of mortality among mechanically ventilated patients in the intensive care units of ethiopian hospitals and the associated factors: A systematic review and meta-analysis. *PLOS ONE*, 19(7):e0306277, July 2024.
- [5] Nicolas Paul, Elena Ribet Buse, Julius J. Grunow, Stefan J. Schaller, Claudia D. Spies, Andreas Edel, and Björn Weiss. Prolonged mechanical ventilation in critically ill patients: Six-month mortality, care pathways, and quality of life. *CHEST*, 168(1):106–118, July 2025.
- [6] Joseph F. Dasta, Trent P. McLaughlin, Samir H. Mody, and Catherine Tak Piech. Daily cost of an intensive care unit day: the contribution of mechanical ventilation. *Critical Care Medicine*, 33(6):1266–1271, June 2005.
- [7] Klaus Kaier, Thomas Heister, Edith Motschall, Philip Hehn, Tobias Bluhmki, and Martin Wolkewitz. Impact of mechanical ventilation on the daily costs of icu care: a systematic review and meta regression. *Epidemiology and Infection*, 147:e314, December 2019.
- [8] Klaus Kaier, Thomas Heister, Jan Wolff, and Martin Wolkewitz. Mechanical ventilation and the daily cost of icu care. *BMC Health Services Research*, 20:267, March 2020.

- [9] Belen Cabello and Jordi Mancebo. Work of breathing. *Intensive Care Medicine*, 32(9):1311–1314, September 2006.
- [10] John N. Cronin, Luigi Camporota, and Federico Formenti. Mechanical ventilation in covid-19: A physiological perspective. *Experimental Physiology*, 107(7):683–693, July 2022.
- [11] Tàì Pham, Laurent J. Brochard, and Arthur S. Slutsky. Mechanical ventilation: State of the art. *Mayo Clinic Proceedings*, 92(9):1382–1400, September 2017.
- [12] Bhavesh Popat and Andrew T. Jones. Invasive and non-invasive mechanical ventilation. *Medicine*, 44(6):346–350, June 2016.
- [13] Martin Dres, Bruno-Pierre Dubé, Julien Mayaux, Julie Delemazure, Danielle Reuter, Laurent Brochard, Thomas Similowski, and Alexandre Demoule. Coexistence and impact of limb muscle and diaphragm weakness at time of liberation from mechanical ventilation in medical intensive care unit patients. *American Journal of Respiratory and Critical Care Medicine*, 195(1):57–66, January 2017.
- [14] Ewan C. Goligher, Eddy Fan, Margaret S. Herridge, Alistair Murray, Stefannie Vorona, Debbie Brace, Nuttapol Rittayamai, Ashley Lanys, George Tomlinson, Jeffrey M. Singh, Steffen-Sebastian Bolz, Gordon D. Rubenfeld, Brian P. Kavanagh, Laurent J. Brochard, and Niall D. Ferguson. Evolution of diaphragm thickness during mechanical ventilation. impact of inspiratory effort. *American Journal of Respiratory and Critical Care Medicine*, 192(9):1080–1088, November 2015.
- [15] Arnaud W. Thille, Pablo Rodriguez, Belen Cabello, François Lellouche, and Laurent Brochard. Patient-ventilator asynchrony during assisted mechanical ventilation. *Intensive Care Medicine*, 32(10):1515–1522, October 2006.
- [16] Lluís Blanch, Ana Villagra, Bernat Sales, Jaume Montanya, Umberto Lucangelo, Manel Luján, Oscar García-Esquirol, Encarna Chacón, Anna Estruga, Joan C. Oliva, Alberto Hernández-Abadia, Guillermo M. Albaiceta, Enrique Fernández-Mondejar, Rafael Fernández, Josefina Lopez-Aguilar, Jesús Villar, Gastón Murias, and Robert M. Kacmarek. Asynchronies during mechanical ventilation are associated with mortality. *Intensive Care Medicine*, 41(4):633–641, April 2015.
- [17] E. Wesley Ely, Sharon K. Inouye, Gordon R. Bernard, Sharon Gordon, Joseph Francis, Lisa May, Brenda Truman, Theodore Speroff, Shiva Gautam, Richard Margolin, Robert P. Hart, and Robert Dittus. Delirium in mechanically ventilated patients: Validity and reliability of the confusion assessment method for the intensive care unit (cam-icu). *JAMA*, 286(21):2703–2710, December 2001.
- [18] M. Chance Spalding, Michael W. Cripps, and Christian T. Minshall. Ventilator-associated pneumonia: New definitions. *Critical Care Clinics*, 33(2):277–292, April 2017.
- [19] Fiona Howroyd, Cyril Chacko, Andrew MacDuff, Nandan Gautam, Brian Pouchet, Bill Tunnicliffe, Jonathan Weblin, Fang Gao-Smith, Zubair Ahmed, Niharika A. Duggal, and Tonny Veenith. Ventilator-associated pneumonia: pathobiological heterogeneity and diagnostic challenges. *Nature Communications*, 15:6447, July 2024.
- [20] Stephan von Düring, Ken Kuljit S. Parhar, Neill K. J. Adhikari, Martin Urner, S. Joseph Kim, Laveena Munshi, Kuan Liu, and Eddy Fan. Understanding ventilator-induced lung injury: The role of mechanical power. *Journal of Critical Care*, 85:154902, February 2025.

- [21] M. R. Pinsky. The hemodynamic consequences of mechanical ventilation: an evolving story. *Intensive Care Medicine*, 23(5):493–503, May 1997.
- [22] Eyob Eshete Tadesse, Ambaye Dejen Tilahun, Nurhusein Nuru Yesuf, Teshome Demis Nimani, and Tesfaye Ayenew Mekuria. Mortality and its associated factors among mechanically ventilated adult patients in the intensive care units of referral hospitals in northwest amhara, ethiopia, 2023. *Frontiers in Medicine*, 11:1345468, July 2024.
- [23] Nadir Yehya, Michael O. Harhay, Martha A. Q. Curley, David A. Schoenfeld, and Ron W. Reeder. Reappraisal of ventilator-free days in critical care research. *American Journal of Respiratory and Critical Care Medicine*, 200(7):828–836, October 2019.
- [24] Annachiara Marra, Ryan J. Smith, Pratik P. Pandharipande, and Christina J. Hayhurst. Optimal sedation in ventilated patients with sepsis and septic shock. *Critical Care Clinics*, 41(4):775–790, October 2025.
- [25] Patel B. Shruti and John P. Kress. Sedation and analgesia in the mechanically ventilated patient. *American journal of respiratory and critical care medicine*, 185(5):486–497, March 2012.
- [26] José Miguel Marcos-Vidal, Rafael González, María Merino, Eva Higuera, and Cristina García. Sedation for patients with sepsis: Towards a personalised approach. *Journal of Personalized Medicine*, 13(12):1641, November 2023.
- [27] Preet Mohinder Singh, Vimi Rewari, Chandralekha, Mahesh Kumar Arora, and Anjan Trikha. A retrospective analysis of determinants of self-extubation in a tertiary care intensive care unit. *Journal of Emergencies, Trauma, and Shock*, 6(4):241–245, October 2013.
- [28] Fabio Tanzillo Moreira and Ary Serpa Neto. Sedation in mechanically ventilated patients—time to stay awake? *Annals of Translational Medicine*, 4(19):382, October 2016.
- [29] Valerie Page and Cathy McKenzie. Sedation in the intensive care unit. *Current Anesthesiology Reports*, 11(2):92–100, April 2021.
- [30] Leanne M. Aitken, Kalliopi Kydonaki, Bronagh Blackwood, Laurence G. Trahair, Edward Purssell, Mandeep Sekhon, and Timothy S. Walsh. Inconsistent relationship between depth of sedation and intensive care outcome: systematic review and meta-analysis. *Thorax*, 76(11):1089–1098, November 2021.
- [31] Nicola Gitti, Stefania Renzi, Mattia Marchesi, Michele Bertoni, Francisco A. Lobo, Frank A. Rasulo, Alberto Goffi, Matteo Pozzi, and Simone Piva. Seeking the light in intensive care unit sedation: The optimal sedation strategy for critically ill patients. *Frontiers in Medicine*, 9:901343, June 2022.
- [32] Robert J. Stephens, Erin M. Evans, Michael J. Pajor, Ryan D. Pappal, Haley M. Egan, Max Wei, Hunter Hayes, Jason A. Morris, Nicholas Becker, Brian W. Roberts, Marin H. Kollef, Nicholas M. Mohr, and Brian M. Fuller. A dual-center cohort study on the association between early deep sedation and clinical outcomes in mechanically ventilated patients during the covid-19 pandemic: The covid-sed study. *Critical Care*, 26(1):179, June 2022.

- [33] Robert J. Stephens, Matthew R. Dettmer, Brian W. Roberts, Enyo Ablordeppey, Susan A. Fowler, Marin H. Kollef, and Brian M. Fuller. Practice patterns and outcomes associated with early sedation depth in mechanically ventilated patients: A systematic review and meta-analysis. *Critical Care Medicine*, 46(3):471–479, March 2018.
- [34] Jiayue Xu, Qiao He, Mingqi Wang, Zichen Wang, Wenkai Wu, Lingling Li, Wen Wang, and Xin Sun. Early deep-to-light sedation versus continuous light sedation for icu patients with mechanical ventilation: A cohort study. *Anaesthesia Critical Care Pain Medicine*, 43(6):101441, December 2024.
- [35] Lilian Maria Sobreira Tanaka, Rodrigo Bernardo Serafim, and Jorge Ibrain Figueira Sal-luh. What every intensivist should know about light sedation for mechanically ventilated patients. *Revista Brasileira de Terapia Intensiva*, 33(4):480–482, November 2021.
- [36] Kimberley Lewis, Michele C. Balas, Joanna L. Stollings, Molly McNett, Timothy D. Girard, Gerald Chanques, Michelle E. Kho, Pratik P. Pandharipande, Gerald L. Weinhouse, Nathan E. Brummel, Linda L. Chlan, Makayla Cordoza, Jeremiah J. Duby, Céline Gélinas, Erin L. Hall-Melnchuk, Anna Krupp, Patricia R. Louzon, Judith A. Tate, Bethany Young, Ron Jennings, Anitra Hines, Chris Ross, Kallirroi Laiya Carayannopoulos, and J. Matthew Aldrich. A focused update to the clinical practice guidelines for the prevention and management of pain, anxiety, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the icu. *Critical Care Medicine*, 53(3):e711–e727, March 2025.
- [37] Curtis N. Sessler, Mark S. Gosnell, Mary Jo Grap, Gretchen M. Brophy, Pam V. O’Neal, Kimberly A. Keane, Eljim P. Tesoro, and R.K. Elswick. The richmond agitation-sedation scale: validity and reliability in adult intensive care unit patients. *American journal of respiratory and critical care medicine*, 166(10):1338–1344, November 2002.
- [38] Marko M. Sahinovic, Michel M. R. F. Struys, and Anthony R. Absalom. Clinical pharmacokinetics and pharmacodynamics of propofol. *Clinical Pharmacokinetics*, 57(12):1539–1558, July 2018.
- [39] Christopher G. Hughes, Patrick T. Mailloux, John W. Devlin, Joshua T. Swan, Robert D. Sanders, Antonio Anzueto, James C. Jackson, Aimee S. Hoskins, Brenda T. Pun, Onur M. Orun, Rameela Raman, Joanna L. Stollings, Amy L. Kiehl, Matthew S. Duprey, Lan N. Bui, Hollis R. O’Neal, Allison Snyder, Michael A. Gropper, Kalpalatha K. Guntupalli, Gregg J. Stashenko, Mayur B. Patel, Nathan E. Brummel, Timothy D. Girard, Robert S. Dittus, Gordon R. Bernard, E. Wesley Ely, and Pratik P. Pandharipande. Dexmedetomidine or propofol for sedation in mechanically ventilated adults with sepsis. *New England Journal of Medicine*, 384(15):1424–1436, February 2021.
- [40] G. Trapani, C. Altomare, E. Sanna, G. Biggio, and G. Liso. Propofol in anesthesia. mechanism of action, structure-activity relationships, and drug delivery. *Current Medicinal Chemistry*, 7(2):249–271, February 2000.
- [41] Jananthan Paramsothy, Sai Dheeraj Gutlapalli, Vijay Durga Pradeep Ganipineni, Isabelle Mulango, Ikpechukwu J Okorie, Divine Besong Arrey Agbor, Crystal Delp, Hanim Apple, Borislav Kheyson, Jay Nfonoyim, Nidal Isber, and Mallikarjuna Yalamanchili. Propofol in icu settings: Understanding and managing anti-arrhythmic, pro-arrhythmic effects, and propofol infusion syndrome. *Cureus*, 15(6):e40456, June 2023.

- [42] Staffan Lundström, Robert Twycross, Mary Mihalyo, and Andrew Wilcock. Propofol. *Journal of Pain and Symptom Management*, 40(3):466–470, September 2010.
- [43] Yijun Seo, Hak-Jae Lee, Eun Jin Ha, and Tae Sun Ha. 2021 ksccm clinical practice guidelines for pain, agitation, delirium, immobility, and sleep disturbance in the intensive care unit. *Acute and Critical Care*, 37(1):1–25, February 2022.
- [44] Aayushi Singh and Ashish P Anjankar. Propofol-related infusion syndrome: A clinical review. *Cureus*, 14(10):e30383, October 2022.
- [45] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, February 2016.
- [46] Jonathan H. Chen and Steven M. Asch. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *The New England Journal of Medicine*, 376(26):2507–2509, June 2017.
- [47] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, July 2020.
- [48] Miguel A. Hernán, John Hsu, and Brian Healy. A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, 32(1):42–49, March 2019.
- [49] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, New York, May 2018.
- [50] Aleksander Molak. *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more*. Packt Publishing, Birmingham, 1 edition, 2023.
- [51] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, December 1995.
- [52] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 53(4):1–37, July 2021.
- [53] Daniel Rodriguez-Cardenas, Aya Garryeva, David N. Palacio, Antonio Mastropaolo, and Denys Poshyvanyk. Rethinking software empirical studies with structural causal models. *arXiv preprint arXiv:2605.28482*, May 2026.
- [54] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, September 1988.
- [55] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, June 2019.

- [56] Peter Spirtes. *Causation, prediction, and search*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2nd ed edition, 2000.
- [57] Bryan Andrews, Joseph Ramsey, Rubén Sánchez-Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of dags using the best order score search and grow-shrink trees. *arXiv preprint arXiv:2310.17679*, October 2023.
- [58] Martin Huber. An introduction to causal discovery. *Swiss Journal of Economics and Statistics*, 160:14, October 2024.
- [59] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2 edition, 2009.
- [60] Juan Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.
- [61] Drago Plečko, Domagoj Bradač, Matija Bucić, and Elias Bareinboim. Beyond the backdoor: Probabilities of identification. Technical Report R-125, Columbia University, May 2025.
- [62] Bryan Andrews, Joseph Ramsey, and Gregory F. Cooper. Learning high-dimensional directed acyclic graphs with mixed data-types. In *Proceedings of Machine Learning Research*, volume 104, pages 4–21. PMLR, August 2019.
- [63] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3921–3962, November 2014.
- [64] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, November 2002.
- [65] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, March 2017.
- [66] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512, 1986.
- [67] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.
- [68] Lan Wen, Jessica G. Young, James M. Robins, and Miguel A. Hernán. Parametric g-formula implementations for causal survival analyses. *Biometrics*, 77(2):740–753, June 2021.
- [69] CAUSALab. pygformula documentation: Algorithm outline, 2024.
- [70] Ashley I. Naimi, Stephen R. Cole, and Edward H. Kennedy. An introduction to g methods. *International Journal of Epidemiology*, 46(2):756–762, April 2017.
- [71] Jacqueline E. Rudolph, Abigail Cartus, Lisa M. Bodnar, Enrique F. Schisterman, and Ashley I. Naimi. The role of the natural course in causal analysis. *American Journal of Epidemiology*, 191(2):341–348, February 2022.

- [72] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, November 2020.
- [73] Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, August 2017.
- [74] Jingpu Shi and Beau Norgeot. Learning causal effects from observational data in healthcare: A review and summary. *Frontiers in Medicine*, 9, July 2022.
- [75] J. M. Smit, J. H. Krijthe, W. M. R. Kant, J. A. Labrecque, M. Komorowski, D. a. M. P. J. Gommers, J. van Bommel, M. J. T. Reinders, and M. E. van Genderen. Causal inference using observational intensive care unit data: a scoping review and recommendations for future practice. *npj Digital Medicine*, 6:221, November 2023.
- [76] Shuai Zhao, Ruihui Zhou, Qi Zhong, and Mi Zhang. Effect of age and icu types on mortality in invasive mechanically ventilated patients with sepsis receiving dexmedetomidine: a retrospective cohort study with propensity score matching. *Frontiers in Pharmacology*, 15, February 2024.
- [77] Gyungah Kim, Dong Kyu Oh, Su Yeon Lee, Mi Hyeon Park, and Chae-Man on behalf of the Korean Sepsis Alliance (KSA) investigators Lim. Impact of the timing of invasive mechanical ventilation in patients with sepsis: a multicenter cohort study. *Critical Care*, 28:297, September 2024.
- [78] Junghwan Lee, Simin Ma, Nicoleta Serban, and Shihao Yang. Accurate treatment effect estimation using inverse probability of treatment weighting with deep learning. *JAMIA Open*, 8(2):ooaf032, April 2025.
- [79] João Matos, Tristan Struja, Naira Link Woite, David Restrepo, Andre Kurepa Waschka, Leo A. Celi, and Christopher M. Sauer. A causal inference framework to compare the effectiveness of life-sustaining icu therapies—using the example of cancer patients with sepsis. *International Journal of Cancer*, 158(3):707–715, February 2026.
- [80] Haoyang Li, Chengxi Zang, Zhenxing Xu, Weishen Pan, Suraj Rajendran, Yong Chen, and Fei Wang. Federated target trial emulation using distributed observational data for treatment effect estimation. *npj Digital Medicine*, 8:387, July 2025.
- [81] Changchang Yin, Hong-You Chen, Wei-Lun Chao, and Ping Zhang. Federated inverse probability treatment weighting for individual treatment effect estimation. *ACM Transactions on Intelligent Systems and technology*, 17(2):1–26, February 2026.
- [82] Md Osman Gani, Shravan Kethireddy, Riddhiman Adib, Uzma Hasan, Paul Griffin, and Mohammad Adibuzzaman. Structural causal model with expert augmented knowledge to estimate the effect of oxygen therapy on mortality in the icu. *Artificial Intelligence in Medicine*, 137:102493, March 2023.
- [83] TEAM Study Investigators, the ANZICS Clinical Trials Group, Carol L. Hodgson, Michael Bailey, Rinaldo Bellomo, Kathy Brickell, Tessa Broadley, Heidi Buhr, Belinda J. Gabbe, Doug W. Gould, Meg Harrold, Alisa M. Higgins, Sally Hurford, Theodore J. Iwashyna, Ary Serpa Neto, Alistair D. Nichol, Jeffrey J. Presneill, Stefan J. Schaller, Janani Sivasuthan, Claire J. Tipping, Steven Webb, and Paul J. Young. Early active mobilization during mechanical ventilation in the icu. *The New England Journal of Medicine*, 387(19):1747–1758, 2022.

- [84] Yahya Shehabi, Belinda D. Howe, Rinaldo Bellomo, Yaseen M. Arabi, Michael Bailey, Frances E. Bass, Suhaini Bin Kadiman, Colin J. McArthur, Lynnette Murray, Michael C. Reade, Ian M. Seppelt, Jukka Takala, Matt P. Wise, and for the SPICE III Investigators Webb, Steven A. for the ANZICS Clinical Trials Group. Early sedation with dexmedetomidine in critically ill patients. *New England Journal of Medicine*, 380(26):2506–2517, May 2019.
- [85] Kiyan Heybati, Fangwen Zhou, Saif Ali, Jiawen Deng, Divyanshu Mohananey, Pedro Villablanca, and Harish Ramakrishna. Outcomes of dexmedetomidine versus propofol sedation in critically ill adults requiring mechanical ventilation: a systematic review and meta-analysis of randomised controlled trials. *British Journal of Anaesthesia*, 129(4):515–526, October 2022.
- [86] Longxiang Su, Chun Liu, Fengxiang Chang, Bo Tang, Lin Han, Huizhen Jiang, Weiguo Zhu, Na Hong, Xiang Zhou, and Yun Long. Selection strategy for sedation depth in critically ill patients on mechanical ventilation. *BMC Medical Informatics and Decision Making*, 21:79, July 2021.
- [87] Laurent Papazian, Jean-Marie Forel, Arnaud Gacouin, Christine Penot-Ragon, Gilles Perrin, Anderson Loundou, Samir Jaber, Jean-Michel Arnal, Didier Perez, Jean-Marie Seghboyan, Jean-Michel Constantin, Pierre Courant, Jean-Yves Lefrant, Claude Guérin, Gwenaël Prat, Sophie Morange, and Antoine Roch. Neuromuscular blockers in early acute respiratory distress syndrome. *New England Journal of Medicine*, 363(12):1107–1116, September 2010.
- [88] Hanne T. Olsen, Helene K. Nedergaard, Thomas Strøm, Jakob Oxlund, Karl-Andre Wian, Lars M. Ytrebø, Bjørn A. Kroken, Michelle Chew, Serkan Korkmaz, Jørgen T. Lauridsen, and Palle Toft. Nonsedation or light sedation in critically ill, mechanically ventilated patients. *New England Journal of Medicine*, 382(12):1103–1111, February 2020.
- [89] Niloufar Eghbali, Tuka Alhanai, and Mohammad M. Ghassemi. Patient-specific sedation management via deep reinforcement learning. *Frontiers in Digital Health*, 3, March 2021.
- [90] Niloufar Eghbali, Tuka Alhanai, and Mohammad M. Ghassemi. Reinforcement learning approach to sedation and delirium management in the intensive care unit. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5, 2023.
- [91] Hong Yeul Lee, Soomin Chung, Dongwoo Hyeon, Hyun-Lim Yang, Hyung-Chul Lee, Ho Geol Ryu, and Hyeonhoon Lee. Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients. *npj Digital Medicine*, 7:325, November 2024.
- [92] Joel Romero-Hernandez and Oscar Camara. On safer reinforcement learning for sedation and analgesia in intensive care. *arXiv preprint arXiv:2601.23154*, May 2026.
- [93] Daniel Lopez-Martinez, Patrick Eschenfeldt, Sassan Ostvar, Myles Ingram, Chin Hur, and Rosalind Picard. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep q networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, page 3960–3963, July 2019.

- [94] Alfonso Muriel, Oscar Peñuelas, Fernando Frutos-Vivar, Alejandro C. Arroliga, Victor Abaira, Arnaud W. Thille, Laurent Brochard, Nicolás Nin, Andrew R. Davies, Pravin Amin, Bin Du, Konstantinos Raymonds, Fernando Rios, Damian A. Violi, Salvatore M. Maggiore, Marco Antonio Soares, Marco González, Fekri Abroug, Hans-Henrik Bülow, Javier Hurtado, Michael A. Kuiper, Rui P. Moreno, Amine Ali Zeggwagh, Asisclo J. Villagómez, Manuel Jibaja, Luis Soto, Gabriel D’Empaire, Dimitrios Matamis, Younsuck Koh, Antonio Anzueto, Niall D. Ferguson, and Andrés Esteban. Impact of sedation and analgesia during noninvasive positive pressure ventilation on outcome: a marginal structural model causal analysis. *Intensive Care Medicine*, 41(9):1586–1600, September 2015.
- [95] Peng-Yen Wu, Shang-Yi Lee, Li-Ting Wong, and Wen-Cheng Chao. Early deep sedation was associated with post-hospital one-year mortality in critically ill surgical patients: a propensity-matched retrospective cohort study. *BMC Anesthesiology*, 25:268, May 2025.
- [96] Eric D. Sinzinger and Brett Moore. Sedation of simulated icu patients using reinforcement learning based control. *International Journal on Artificial Intelligence Tools*, 14(01n02):137–156, 2005.
- [97] Regina Padmanabhan, Nader Meskin, and Wassim M. Haddad. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomedical Signal Processing and Control*, 22:54–64, September 2015.
- [98] Regina Padmanabhan, Nader Meskin, and Wassim M. Haddad. Reinforcement learning-based control for combined infusion of sedatives and analgesics. In *2017 4th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 0505–0509, April 2017.
- [99] Benjamin Ribba, Dominic Stefan Bräm, Paul Gabriel Baverel, and Richard Wilson Peck. Model enhanced reinforcement learning to enable precision dosing: A theoretical case study with dosing of propofol. *CPT: Pharmacometrics & Systems Pharmacology*, 11(11):1497–1510, November 2022.
- [100] Junqi Cui, Weijia Li, Enoch Chi Ngai Lim, Xiaoqin Wu, and Chi Eung Danforn Lim. Stratified causal inference for intensive care unit risk prediction: Informatics-based modeling of anesthetic drug combinations. *JMIR Formative Research*, 10(1):e80294, February 2026.
- [101] Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Position: reinforcement learning in dynamic treatment regimes needs critical reexamination. *arXiv preprint arXiv:2405.18556*, May 2024.
- [102] Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5942–5962, April 2025.
- [103] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.
- [104] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. Version 3.1.

- [105] Hong Liu, Fuhai Ji, Ke Peng, Richard L. Applegate, and Neal Fleming. Sedation after cardiac surgery: Is one drug better than another? *Anesthesia Analgesia*, 124(4):1061–1070, April 2017.
- [106] Stéphane Bertin, David Haefliger, Antoine G. Schneider, Raphaël Giraud, Maria-Helena Perez, Xavier Bechtold, Ermindo R. Di Paolo, Laura E. Rothuizen, Thierry Buclin, and Françoise Livio. Effects of extracorporeal membrane oxygenation circuits on drug sequestration: A review of ex vivo experiments. *Journal of Clinical Medicine*, 14(22):8060, November 2025.
- [107] Wanjun Liu, Gan Tao, Yijun Zhang, Wenyan Xiao, Jin Zhang, Yu Liu, Zongqing Lu, Tianfeng Hua, and Min Yang. A simple weaning model based on interpretable machine learning algorithm for patients with sepsis: A research of mimic-iv and eicu databases. *Frontiers in Medicine*, 8:814566, January 2022.
- [108] Pallavi Deulkar, Amol Singam, V. N. K. Srinivas Mudiganti, and Abhishek Jain. Lactate monitoring in intensive care: A comprehensive review of its utility and interpretation. *Cureus*, 16(8):e66356, August 2024.
- [109] Jill Vanmassenhove, Norbert Lameire, Annemieke Dhondt, Raymond Vanholder, and Wim Van Biesen. Prognostic robustness of serum creatinine based aki definitions in patients with sepsis: a prospective cohort study. *BMC Nephrology*, 16:112, July 2015.
- [110] John A. Kellum and Norbert Lameire. Diagnosis, evaluation, and management of acute kidney injury: a kdigo summary (part 1). *Critical Care*, 17(1):204, February 2013.
- [111] Peter C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107, 2009.
- [112] Joseph Ramsey and Bryan Andrews. Py-tetrad and rpy-tetrad: A new python interface with r support for tetrad causal search. In *Proceedings of the 2023 Causal Analysis Workshop Series*, volume 223, pages 40–51. PMLR, August 2023.
- [113] Ryan M. Andrews, Ronja Foraita, Vanessa Didelez, and Janine Witte. A practical guide to causal discovery with cohort data. *arXiv preprint arXiv:2108.13395*, December 2023.
- [114] CAUSALab. pygformula: A python implementation of the parametric g-formula, 2024.
- [115] CAUSALab. pygformula documentation: Output, 2024.
- [116] You-Qun Huang, Rong Gou, Yong-Shu Diao, Qing-Hua Yin, Wen-Xing Fan, Ya-Ping Liang, Yi Chen, Min Wu, Li Zang, Ling Li, Jing Zang, Lu Cheng, Ping Fu, and Fang Liu. Charlson comorbidity index helps predict the risk of mortality for patients with type 2 diabetic nephropathy. *Journal of Zhejiang University SCIENCE B*, 15:58–66, January 2014.
- [117] Mahmoud A. Ammar, Abdalla A. Ammar, Patrick M. Wieruszewski, Brittany D. Bissell, Micah T. Long, Lauren Albert, Ashish K. Khanna, and Gretchen L. Sacha. Timing of vasoactive agents and corticosteroid initiation in septic shock. *Annals of Intensive Care*, 12(1):47, May 2022.

- [118] Ary Serpa Neto, Marcus J. Schultz, and Emir Festic. Ventilatory support of patients with sepsis or septic shock in resource-limited settings. *Intensive Care Medicine*, 42:100–103, January 2016.
- [119] Thomas W. Schnider, Charles F. Minto, Steven L. Shafer, Pedro L. Gambus, Corina Andresen, David B. Goodale, and Elizabeth J. Youngs. The influence of age on propofol pharmacodynamics. *Anesthesiology*, 90(6):1502–1516, June 1999.
- [120] Yi Lin Lee, Kalyanasundaram Ganesh, Lian Kah Ti, and Shin Yi Ng. A prospective, observational, longitudinal cohort study of sedation practices in sgh intensive care units. *Proceedings of Singapore Healthcare*, 27(2):103–109, September 2018.
- [121] Yu-Han Chiu, Lan Wen, Sean McGrath, Roger Logan, Issa J. Dahabreh, and Miguel A. Hernán. Evaluating model specification when using the parametric g-formula in the presence of censoring. *American Journal of Epidemiology*, 192(11):1887–1895, November 2023.
- [122] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), November 2019.
- [123] Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *arXiv preprint arXiv:2007.01884*, February 2021.