An Algorithmic Approach for Causal Health Equity: A Look at Race Differentials in Intensive Care Unit (ICU) Outcomes

Drago Plečko[1*], Paul Secombe[2], Andrea Clarke[3], Amelia Fiske[4], Donisha Duff[5], David Pilcher[6], Leo Anthony Celi[7], Rinaldo Bellomo[5], Elias Bareinboim[1]

[1]Department of Computer Science, Columbia University.
[2]School of Medicine, Flinders University.
[3]School of Population and Global Health, University of Melbourne.
[4]Institute for the History and Ethics of Medicine, Technical University of Munich.
[5]Kurongkurl Katitjin, Centre for Indigenous Australian Education and Research, Edith Cowan University.
[6]Australian and New Zealand Intensive Care Research Centre, Monash University.
[7]Laboratory for Computational Physiology, Massachusetts Institute of Technology.

*Corresponding author(s). E-mail(s): dp3144@columbia.edu;
Contributing authors: paulsecombe@bigpond.com ; andrea.clarke@unimelb.edu.au; a.fiske@tum.de; ceo@qibn.com.au; d.pilcher@alfred.org.au; lceli@mit.edu; rinaldo.bellomo@austin.org.au; eb@cs.columbia.edu;

Abstract

Health equity is defined as the state in which everyone has a fair and just opportunity to attain their highest level of health. Achieving health equity is believed to improve the well-being of communities, reduce healthcare costs, and increased productivity and longevity. However, disparities in health are still significant. In this context, the new era of large-scale data collection and analysis presents an opportunity for diagnosing and understanding the causes of health inequities. In this study, we describe a framework for systematically analyzing health disparities using tools of causal inference. We illustrate the framework by investigating racial and ethnic disparities in intensive care unit (ICU) outcome between majority and minority groups in Australia (Indigenous vs. Non-Indigenous) and the United States (African-American vs. White). We demonstrate that commonly used statistical measures for quantifying inequity are insufficient, and focus on attributing the observed disparity to the causal mechanisms that generate it. We find that minority patients are younger at admission, have worse chronic health, are more likely to be admitted for urgent and non-elective reasons, and have higher illness severity. At the same time, however, we also find a protective direct effect of belonging to a minority group, with minority patients showing improved survival compared to their majority counterparts, with all other variables being equal. We then demonstrate that this protective effect is related to the increased probability of being admitted to ICU, with minority patients having an increased risk of ICU admission. Additionally, we also find that minority patients, while showing improved survival, are in fact more likely to be readmitted to ICU. These findings support the hypothesis that, due to worse access to primary health care, minority patients are more likely to end up in ICU for preventable conditions, causing a reduction in the mortality rates and creating an effect that appears to be protective. Since the baseline risk of ICU admission may serve as proxy for lack of access to primary care, we developed the Indigenous Intensive Care Equity (IICE) Radar, a monitoring system for tracking the over-utilization of ICU resources by the Indigenous population of Australia across geographical areas.

Keywords: health equity, causal inference, algorithmic fairness, intensive care medicine

1 Introduction

Health equity is defined as the state in which everyone has a fair and just opportunity to attain their highest level of health (Braveman et al, 2017). One of the key goals in pursuing health equity is the elimination of economic, social, and other obstacles to health and health care, together with eliminating existing, preventable differences in health (Braveman, 2014). The goal is to reduce preventable illnesses and deaths, improve the overall well-being of communities, and create a more inclusive healthcare system that serves everyone effectively (Braveman et al, 2011). Despite a commitment of public health organizations and government agencies to address issues of health equity (Centers for Disease Control and Prevention (CDC), 2024; U.S. Department of Health & Human Services (HHS), 2024b,a; World Health Organization (WHO), 2024; European Union (EU), 2024; Healthy People 2030, 2023), health disparities are still large (Commonwealth Fund, 2023; Dwyer-Lindgren et al, 2022; Health, 2023), and many consider the solutions to health inequities to be in their early stages.

In this context, the rise of the new generation of computational tools and the widespread adoption of electronic health records (EHRs) (Evans, 2016) offer a major opportunity to quantify and understand health disparities. This task is even more important due to a broad transition to using artificial intelligence (AI) tools in healthcare, which may perpetuate or amplify existing biases. In fact, recent works demonstrate formally that understanding disparities in human decision-making is an essential step for understanding disparities in automated AI systems (Plečko and Bareinboim, 2024). Therefore, a successful transition to healthcare systems that benefit from the numerous advantages of AI requires careful scrutiny of the data that is given to such AI systems for training, which includes understanding and quantifying the biases in the data itself. In this paper, we demonstrate that significant disparities in health exist in the data even before applying AI tools. Thus, training AI systems using such biased data may lead to further biases, if not handled appropriately (Plecko and Bareinboim, 2024). At the same time, while great care needs to be taken when training AI systems, a transition to such systems also presents the opportunity to possibly correct for existing human biases.

So far, a systematic framework for analyzing health equity that is compatible with the new needs of the data-driven era has not been proposed. In this paper, we discuss such a framework that may appeal to a broad range of practitioners, data scientists, and AI engineers. The main aim of the framework is to use a causal lens to obtain actionable insights that can improve patient engagement and outcomes. To illustrate this, we perform a case study and focus on the question of racial and ethnic disparities in outcome following admission to an intensive care unit (ICU) (McGowan et al, 2022), in particular focusing on disparities in mortality. We analyze data from Australia (Secombe et al, 2023) for disparities between the First Nations populations of Australia (Aboriginal and Torres Strait Islander) and the majority group (Non-Indigenous population). In parallel, we also analyze a database from a large tertiary hospital in Boston, Massachusetts (Johnson et al, 2020), focusing on the racial disparity between the African-American and White population. We illustrate how a systematic approach for analyzing equity can illuminate different causal mechanisms that generate disparities observed in the data, and demonstrate that racial/ethnic disparities in ICU outcome are markedly similar between the United States and Australia.

2 Results

The research in this paper was approved by the ethics committee of the Alfred Hospital, Melbourne (Project #661/24), and was done in collaboration with the Indigenous Data Network (IDN) of Australia (Indigenous Data Network, 2024).

2.1 Data

The first dataset we analyzed is the Australian and New Zealand (ANZ) Intensive Care Society (ANZICS) Adult Patient Database (APD) (Secombe et al, 2023). We analyzed all admissions from 181 hospitals across Australia between 2018 and 2024. The ANZICS APD receives submissions from 98% of ICUs in Australia.The second dataset we analyzed is the Medical Information Mart for Intensive Care (MIMIC-IV) (Johnson et al, 2020) database, from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. We analyzed all ICU admissions between 2008 and 2019. A detailed discussion of the patient filtering steps, and the description of the covariates used in the analysis, can be found in the Methods section.

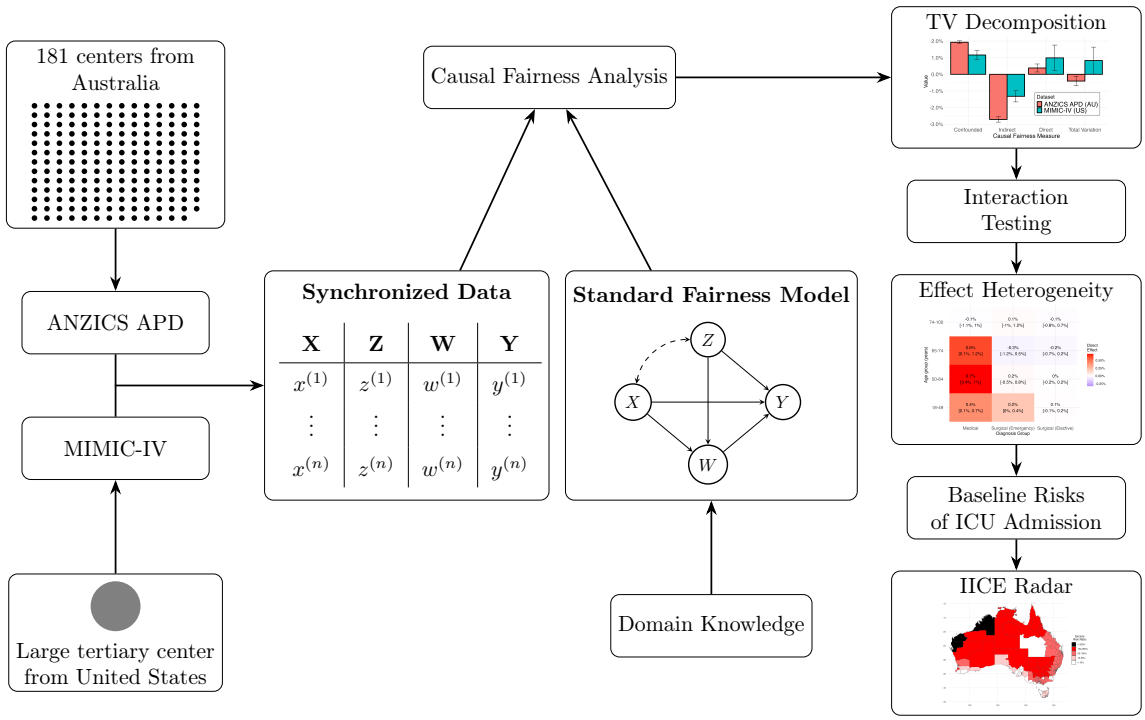


Fig. 1: Diagram of the performed analysis steps.

2.2 Framework of Causal Fairness Analysis

Throughout, we followed the framework of causal fairness analysis described in (Plečko and Bareinboim, 2024). A diagram of the analysis performed in the paper is shown in Fig. 1. Using the ricu R-package (Bennett et al, 2023), we performed data loading for both datasets, which includes information on race (MIMIC-IV) or ethnicity (ANZICS APD), age, sex, socioeconomic status (SES for short, available in ANZICS APD only), illness severity, admission diagnosis, chronic health status (MIMIC-IV only), and in-hospital mortality. Based on domain knowledge, we constructed a specific type of causal diagram called the Standard Fairness Model (SFM), in which variables are categorized into four groups (see Fig. 2a). The protected attribute is race/ethnicity (depending on the dataset), labeled X. The set of confounders, labeled Z, includes age, sex, and SES (SES was available only in Australian data). These demographic variables were chosen as confounders since they may be correlated with race/ethnicity but may not necessarily be causally influenced by it. The set of mediators W includes chronic health status (available in the US data), admission diagnosis, and illness severity. The outcome of interest is in-hospital mortality, labeled Y. A comparison of patient characteristics for majority and minority groups for the ANZICS APD and MIMIC-IV cohorts is given in Tbls. A1 and A2, respectively.

Our goal in this paper was to explain the observed disparity in mortality between the demographic groups. We started by computing the average differences in mortality rates (also known as the total variation, or TV measure), a commonly used measure of disparity, and find that

$$\textbf{(AU)} \quad \mathbb{E}[\text{death} \mid \text{Majority}] - \mathbb{E}[\text{death} \mid \text{First Nations}] = -0.4\%, \tag{1}$$

$$\textbf{(US)} \quad \mathbb{E}[\text{death} \mid \text{White}] - \mathbb{E}[\text{death} \mid \text{African-American}] = 0.8\%. \tag{2}$$

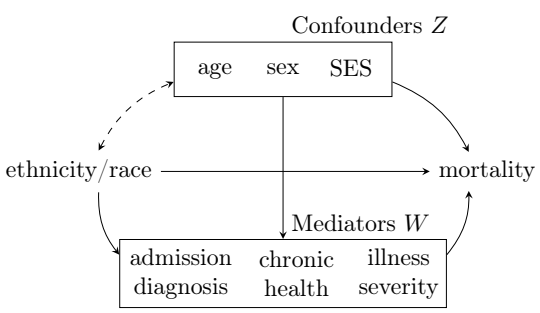
These statistics show that minority patients in Australia were more likely to die after ICU admission, while this finding is reversed in the US data. Based on the constructed causal diagram from Fig. 2a, however, we see that the average difference in mortality rates as in Eqs. 1-2 can arise in three different ways:

(i) *confounded/spurious effect:* race/ethnicity may be associated with age, sex, or socioeconomic status, which may influence the mortality risk,

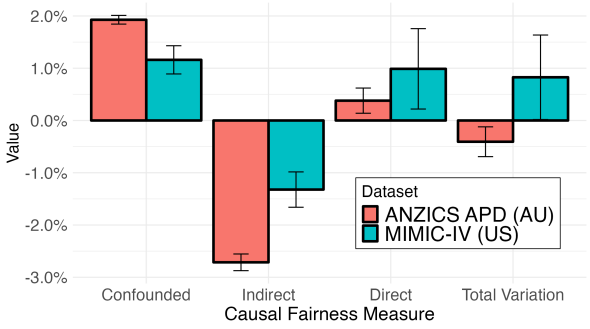
(ii) *indirect effect:* race/ethnicity may influence chronic health, admission diagnosis, and illness severity, which have an effect on the mortality risk,

(iii) *direct effect:* race/ethnicity may influence the mortality risk, with all other variables kept equal.

Using the framework of causal fairness analysis, we can compute the above three effects, and quantify how much each of the effects contributes to the marginal disparity reported in Eqs. 1-2.



(a) Standard Fairness Model.



(b) Causal decompositions ANZICS APD / MIMIC-IV.

Fig. 2: (a) Standard Fairness Model constructed for the data (b) Decomposition of the marginal disparity in outcome into confounded, indirect, and direct effects, on ANZICS APD and MIMIC-IV datasets.

2.3 Decomposing the Disparity

The results quantifying direct, indirect, and confounded effects on both datasets are shown in Fig. 2b. The TV measure, reported in Eqs. 1-2 and shown in the last column of Fig. 2b, takes opposite signs on the two datasets. However, when applying a causal perspective on the problem, we obtain the following decomposition of the TV measure:

$$\textbf{(AU)} \quad \mathbb{E}[\text{death} \mid \text{Majority}] - \mathbb{E}[\text{death} \mid \text{First Nations}] = -0.4\% = \underbrace{1.9\%}_{\text{confounded}} + \underbrace{(-2.7\%)}_{\text{indirect}} + \underbrace{0.4\%}_{\text{direct}}, \quad (3)$$

$$\textbf{(US)} \quad \mathbb{E}[\text{death} \mid \text{White}] - \mathbb{E}[\text{death} \mid \text{African-American}] = 0.8\% = \underbrace{1.1\%}_{\text{confounded}} + \underbrace{(-1.3\%)}_{\text{indirect}} + \underbrace{1.0\%}_{\text{direct}}. \quad (4)$$

We see that the direct, indirect, and spurious effects (first three columns in Fig. 2b) are in fact equal in sign when applying the causal decomposition. The causal interpretation of the decomposition can be summarized as follows. First, along the confounded causal pathway, there exists a protective effect for the First Nations/African-American patients, transmitted through variables such as age, sex, or SES (accounting for 1.9% and 1.1% of overall variation, respectively). Second, along the indirect causal path, there is a harmful effect on minority patients, indicating that minority patients are more likely to die as a result of the indirect effect (accounting for negative 2.7% and 1.3% of the overall variation in AU and US, respectively). Variables that transmit the indirect effect include the admission diagnosis, degree of illness severity, and chronic health status. Third, along the direct causal path (when keeping all other variables equal), there is a protective effect of belonging to the minority group, which accounted for 0.4% of the overall variation in the AU data, and 1.0% in the US data. In other words, for two individuals with comparable characteristics (age, illness severity, admission diagnosis) who differ with respect to the protected attribute, the one belonging to the minority group is more likely to survive after ICU admission. Finally, we emphasize that all the effects studied were statistically significant, indicating that the average differences between groups are in fact a complex interplay of multiple causal mechanisms that transmit change between these demographics.

Explaining the confounded effect. The confounding variables between the protected attribute ethnicity/race and the outcome are age, sex, and SES (see Fig. 2a). We therefore analyzed the age/sex distributions of different subpopulations. In both datasets, minority patients were admitted much younger on average compared to their majority counterparts (average difference in age of 13.6 years in Australian, and 6.3 years in US data). The age distributions for the two countries, across minority and majority groups, are shown in Figs. 3a, 3b. Further, we found differences in socioeconomic status in the Australian data (see Fig. 3c). Indigenous patients have lower SES on average, which increases the mortality risk. Since younger age and higher SES reduce the mortality risk, the group-specific differences in age/SES are both relevant for understanding the confounded effect in the third column of Fig. 2b.

Explaining the indirect effect. The indirect effect, reported in the second column of Fig. 2b, is protective for the majority group. This effect is mediated by admission diagnosis, illness severity, and chronic health status. Therefore, we compared group-specific distributions of illness severity after adjusting for age. In Australia, minority patients had higher illness severity (Fig. 3d, $p < 0.001$ for

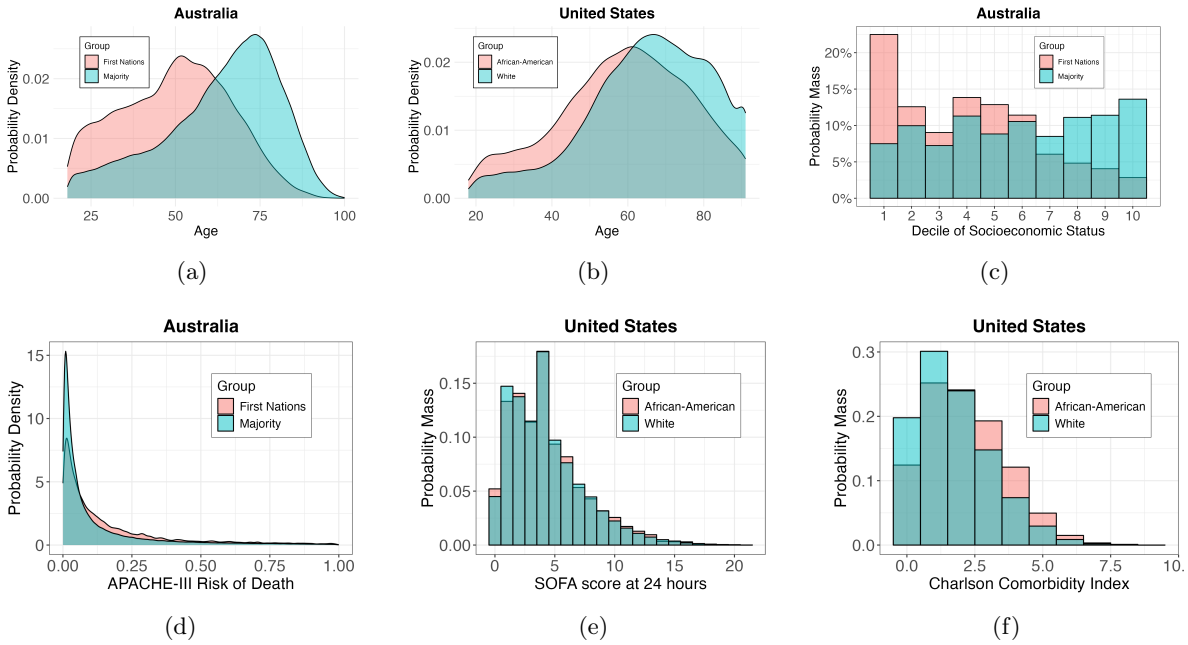


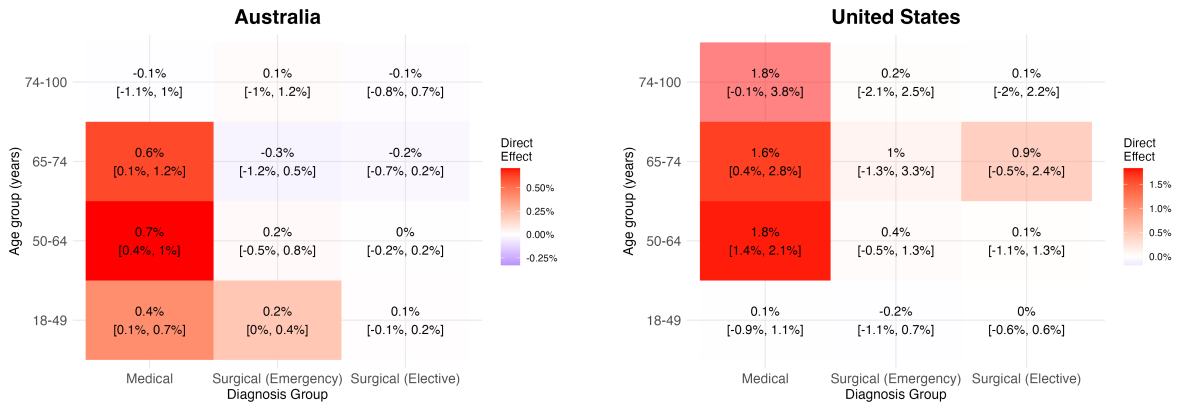
Fig. 3: (a) Age in Australian data; (b) Age in US data; (c) Deciles of socioeconomic status in Australian data; (c) Risk of death (APACHE-III) in Australian data; (d) SOFA score in US data; (e) chronic health status in US data (Charlson comorbidity index).

difference in means), while the distribution of illness severity in the US data did not show a clear trend (Fig. 3e, $p = 0.30$ for difference in means). We also analyzed differential patterns in admission diagnosis after adjusting for age. Minority patients were more likely to be admitted for a medical vs. a surgical reason (both $p < 0.001$) and were more likely to be a non-elective admission (both $p < 0.001$). This mechanism also explains a part of the observed indirect effect, since medical and non-elective admissions carry a greater risk of death. Furthermore, in the US data, where chronic health status is available through the Charlson comorbidity index (Charlson et al, 1987), minority patients had more comorbidities (Fig. 3f, $p < 0.001$) after adjusting for age, also contributing to the indirect effect. In summary, the indirect effect is explained through a combination of differences in admission pattern, chronic health status, and illness severity.

Explaining the direct effect. In both datasets, the direct effect was protective for minority patients when all other variables were kept equal. This finding required an extended analysis, which is described next. First, we noted that possible unobserved confounders for the direct effect of race/ethnicity on outcome include variables such the SES (measured only in AU data) and chronic health status (CHS for short, measured only in US data). However, the above analyses show that SES and CHS (when available) indicate better status for the majority group. Thus, the addition of CHS data to ANZICS APD or the SES data to MIMIC-IV would likely not explain away the protective direct effect for the minority group, but rather make it more pronounced. To better understand direct effects, we took a more granular approach to the quantification of such effects.

2.4 Interaction Testing and Heterogeneous Effects

To understand whether there were important interactions between different causal pathways under study, we performed non-parametric interaction testing (Plecko, 2024). Such methodology provides statistical tests for interactions of causal effects along different pathways, such as the interaction of direct-indirect pathways, or indirect-spurious pathways. Upon applying such an analysis, in both datasets, we found significant interactions of indirect-confounded pathways. In the ANZICS APD data, we additionally found significant interactions of direct-indirect pathways (the full set of p-values for interaction tests is given in Appendix B). The indirect-confounded interaction implies the heterogeneity of indirect effects across different values of the confounder Z, which is studied in Appendix D. The direct-indirect interaction implies heterogeneity of the direct effect along different values of the mediators W. We therefore studied how the direct effect of race/ethnicity on outcome varied across admission diagnoses and age groups. The results of the analysis are shown in Fig. 4, with red color indicating a



(a) Direct effect heterogeneity on ANZICS APD.

(b) Direct effect heterogeneity on MIMIC-IV.

Fig. 4: A heatmap of heterogeneous direct effects of minority status on outcome. The protective effect for the minority groups is driven primarily by medical admissions.

protective direct effect for the minority group. The results indicate that the protective direct effect of minority race/ethnicity was primarily driven by the group of medical admissions in both datasets.

We considered several hypotheses for explaining the protective direct effect for medical admissions of minority patients. Firstly, this effect could represent a genuine biological difference between the majority and minority groups. However, such biological effect would most likely need to be present for both medical and surgical admissions. Further, the fact that we are observing this effect for two entirely distinct populations (Australian First Nations and African-American) that are far apart in terms of ancestral lineages (Tishkoff et al, 2009; Malaspinas et al, 2016) makes this explanation less likely. Therefore, this effect is more likely to be a consequence of social differences between groups, since the minority groups on both continents are known to be disadvantaged in their socioeconomic positions. As discussed earlier, however, it is implausible that confounders such as socioeconomic status or chronic health status would explain away the direct effect that is protective for the minority group.

2.5 Tipping Over Hypothesis and Population Risks

We hypothesized that the protective effect of race/ethnicity observed in the data was due to a *tipping over effect* – minority patients are more likely to require ICU care due to medical complications. The increase in the probability of requiring ICU admission may be a result of worse access to primary health care, since minority groups are less likely to have access to and utilize health care, as reported in previous literature (Davy et al, 2016). Put differently, for majority patients, medical complications are more likely to be prevented through primary care, and therefore only the more severe cases reach the ICU, causing a selection bias (or left-censoring). Following this hypothesis, we sought to investigate the baseline risk of ICU admission across different admission diagnoses in the ANZICS APD. This analysis was possible since the ANZICS APD covers 98% of all ICU admissions in Australia. Data on the age and ethnicity structure of the Australian population were obtained from the Australian Bureau of Statistics (Australian Bureau of Statistics, 2021a,b). A summary of age-adjusted risk ratio for ICU admission between 2018 and 2024, across diagnostic groups, are shown in Fig. 5, with values above 0 indicating that Indigenous patients were more likely to be admitted to the ICU compared to Non-Indigenous patients. The results indicate that the risk of admission for Indigenous patients was uniformly increased for medical and emergency surgery admissions (both of which are urgent, and may be a consequence of suboptimal primary care), while it was mostly decreased for elective surgery admissions (which by definition corresponds to patients participating in primary care). Minority patients were 182% more likely to be admitted to ICU for a medical reason, 79% more likely for emergency surgery, and 14% less likely for elective surgery (all $p < 0.001$).

We sought to investigate if the risk ratio for ICU admission was related to the strength of the observed protective direct effect for the minority group. A possible causal mechanism of the phenomenon may be described as follows. Larger risk ratios for ICU admission indicate that a larger number of patients were admitted for a specific diagnosis. A higher prevalence of a specific diagnosis may decrease the underlying illness severity, since ICU admissions for patients with conditions that may have been prevented at an earlier stage of care are likely to reduce the overall risk in the group. To investigate this potential explanation, we assessed the age-adjusted risk ratio of ICU admission across

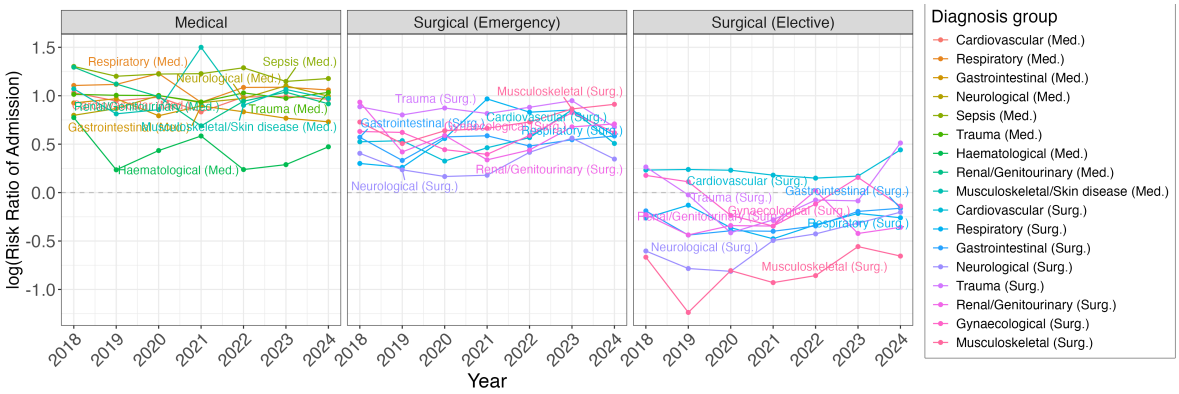
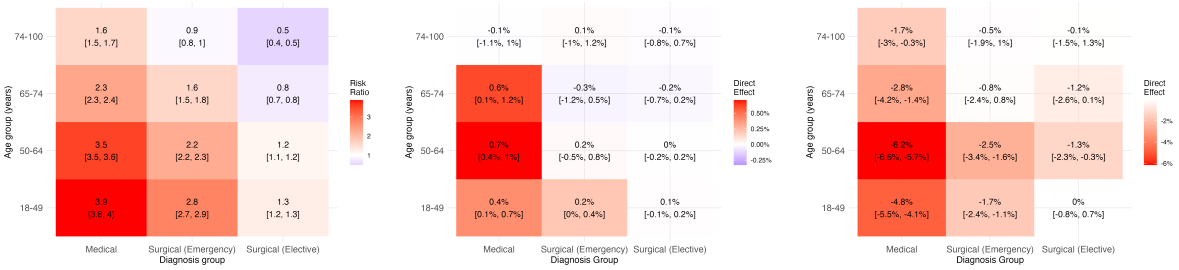


Fig. 5: Logarithm of age-adjusted-$RR(d,t)$ risk ratio (minority group x_0 vs. the baseline majority x_1) for different admission groups over time, between 2018 and 2024.



(a) Baseline risk ratio of ICU admission.

(b) Direct effect of minority status on death.

(c) Direct effect of minority status on readmission.

Fig. 6: Heterogeneity of (a) baseline risk ratio of ICU admission; (b) direct effect of minority status on death; (c) direct effect of minority status on readmission according to age group and admission type.

age groups and admission types, defined as:

$$\frac{P(\text{admission } d \mid do(\text{minority}),\ \text{age group})}{P(\text{admission } d \mid do(\text{majority}),\ \text{age group})}. \tag{5}$$

The larger the ratio, the more likely minority patients are to be admitted for this diagnosis and age group, compared to majority patients. The risk ratios across age groups and admission types are presented in Fig. 6a. Using permutation testing, we tested the similarity of this effect with the estimate direct effect of minority status on mortality and found the similarity to be significant ($p = 0.007$). Therefore, the pattern of increased risk of ICU admission appeared to be related to the protective effect of minority status for different age-diagnosis groups. Logically, however, other unknown factors (such as baseline disease incidence) would provide further insight into this analysis.

2.6 Readmission Analysis

After examining the patterns of protective direct effects and baseline risk of ICU admission, we further investigated the risk of ICU readmission. Within the cohort of patients who survived their hospital stay, we analyzed whether the patient was re-admitted to ICU after being released from the hospital, and this readmission outcome was labeled R. Using the same quantitative approach as in Sec. 2.2, and the same causal model in Fig. 2a with readmission R replacing the mortality outcome Y, we estimated the direct effect of minority status on readmission (along the direct $X \to R$ pathway). The values of this direct effect along different age-admission groups is shown in Fig. 6c, with red color indicating an increased risk of readmission for the minority group. We found this readmission pattern to have significant similarity with both the baseline risk of ICU admission ($p = 0.015$) and the protective direct effect pattern for mortality ($p = 0.02$).

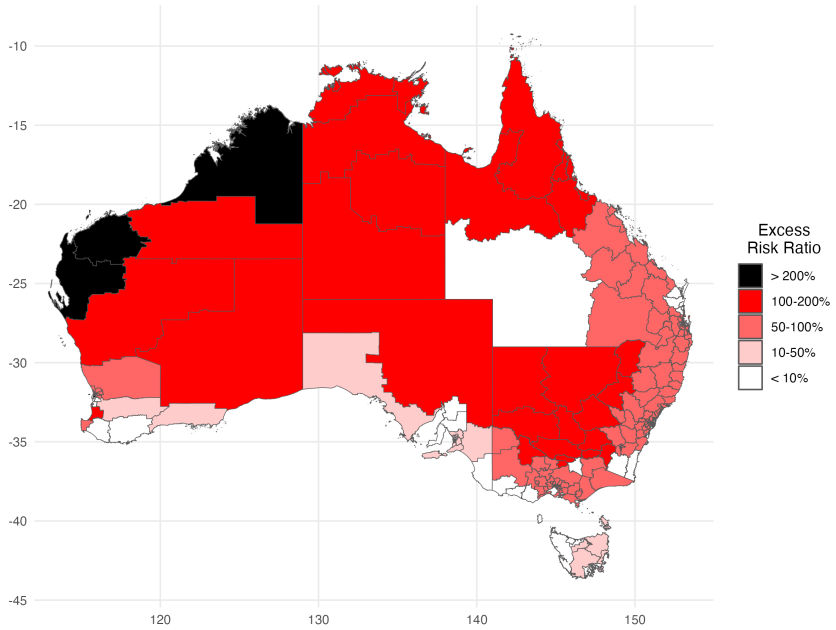


Fig. 7: Indigenous Intensive Care Equity (IICE) Radar.

2.7 Indigenous Intensive Care Equity (IICE) Radar

The fact that Indigenous patients showed improved survival compared to Non-Indigenous patients (with all other variables kept equal) implies that these patients had lower illness severity than the data showed. However, the fact that Indigenous patients were also more likely to be readmitted to ICU, likely implies that the investigated effects may be related to lack of access or under-utilization of primary care. Based on this, we hypothesize that the increase in the baseline risk of ICU admission is a proxy for the under-utilization of primary care by Indigenous patients. This motivated the construction of the Indigenous Intensive Care Equity (IICE) Radar. From the Australian Bureau of Statistics (Australian Bureau of Statistics, 2021b), we extracted data about age and Indigenous status across different geographical areas of Australia. We then matched areas with the patient information available in the ANZICS APD, to compute the age-adjusted baseline risk ratio of ICU admission, i.e., how much more likely minority patients were to be admitted to ICU depending on the area (our analysis was performed on the level of Statistical Areas 3 used by the Australian Bureau of Statistics for the year 2021). The IICE Radar is shown in Fig. 7, with five risk categories for the different areas. Excess risk ratio of less than 10% was considered as no risk (white color). Excess RR of 10-50% was classified as moderate risk (light pink), 50-100% as substantial (light red), 100-200% as severe (red), and >200% as extreme (black). Many of the statistical areas experienced at least moderate risk, implying Indigenous patients were 50% to a 100% more likely to be admitted to ICU. Some parts of Western Australia (South West - Perth) recorded excess risk ratios of more than 300% (minority patients were four times more likely to be admitted), while in other parts (Outback North) the excess risk was up to 900% (minority patients were ten times more likely to be admitted to ICU). The IICE Radar uncovers substantial differences in risk of ICU admission across different regions, requiring further investigation. The construction of the Radar opens the door to monitoring the increase in admission to ICU, and future studies of access to primary care and its impact on health outcomes of minority groups.

3 Methods

3.1 Dataset Description and Covariates

The ANZICS APD contains information on demographics (including age, sex, First Nations status), biochemical and physiological values during the first 24 hours of the ICU stay, and ICU admission diagnosis (ANZICS modification of the APACHE IV diagnosis list). The information is used for the calculation of severity of illness scores including Sequential Organ Failure Assessment (SOFA), Acute Physiology and Chronic Health Evaluation (APACHE) III scores, APACHE III predicted risk of death, and in-hospital outcomes. The database also contains postcode information for each patient, and the postcode can be matched to the Index of Relative Socioeconomic Advantage and Disadvantage (IRSAD)

developed by the Australian Bureau of Statistics (Australian Bureau of Statistics, 2021c), providing information on socioeconomic status (SES) of patients. The covariates extracted for purposes of analysis were:

- protected attribute X: indigenous status (First Nations and majority group),
- confounders Z: age, sex, and postcode based SES,
- mediators W: the APACHE III (Knaus et al, 1991) predicted risk of death, ANZICS modified APACHE-III admission diagnosis (see full list of considered diagnoses), indicator of whether admission was elective,
- outcome Y: in-hospital mortality.

The MIMIC-IV dataset is a publicly available resource with comprehensive information on patients admitted to the Beth Israel Deaconess Medical Center (BIDMC), in Boston, Massachusetts. It is sourced from two in-hospital database systems, a custom hospital wide EHR and an ICU specific clinical information system. From the dataset, we extracted the following information:

- protected attribute X: race (White and African-American),
- confounders Z: age and sex,
- mediators W: the Sequential Organ Failure Assessment (SOFA) (Vincent et al, 1996) score at 24 hours into ICU stay, the Charlson comorbidity index (Charlson et al, 1987), admission diagnosis from one of the 20 categories available in the MIMIC-IV data (see full list of considered diagnoses),
- outcome Y: in-hospital mortality.

Study flowcharts with patient filtering steps are available in Fig. A1 in Appendix A. Final cohort numbers were $n = 1,035,890$ for ANZICS APD and $n = 38,844$ for MIMIC-IV data.

R Statistical Software Version 4.3.2 was used (R Core Team, 2021) for all the analyses. Data loading was performed using the ricu R-package (Bennett et al, 2023). The README file in our Github repository includes instructions for: (i) installing dependencies; (ii) running a demo analysis requring no data access; (iii) reproducing each figure in the paper; (iv) setting up the MIMIC-IV and ANZICS APD data sources (for which access needs to be obtained).

3.2 Decomposing the Disparity in Outcome

We follow the framework of Causal Fairness Analysis described in (Plečko and Bareinboim, 2024), which is based on the language of structural causal models (SCMs) (Pearl, 2000). Our approach of analyzing and aggregating findings across multiple data sources is also related to the data-fusion paradigm in the causal inference literature (Bareinboim and Pearl, 2016). An SCM is a tuple $\mathcal{M} := \langle V, U, \mathcal{F}, P(u)\rangle$, where V, U are sets of endogenous (observable) and exogenous (latent) variables, respectively, $\mathcal{F}$ is a set of functions f_{V_i}, one for each $V_i \in V$, where $V_i \leftarrow f_{V_i}(\text{pa}(V_i), U_{V_i})$ for some $\text{pa}(V_i) \subseteq V$ and $U_{V_i} \subseteq U$. $P(u)$ is a strictly positive probability measure over U. Each SCM $\mathcal{M}$ is associated to a causal diagram $\mathcal{G}$ (Bareinboim et al, 2022) over the node set V where $V_i \to V_j$ if V_i is an argument of f_{V_j}, and $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding U_{V_i}, U_{V_j} are not independent. An instantiation of the exogenous variables $U = u$ is called a *unit*. By $Y_x(u)$ we denote the potential outcome of Y when setting $X = x$ for the unit u, which is the solution for $Y(u)$ to the set of equations obtained by evaluating the unit u in the submodel $\mathcal{M}_x$, in which all equations in $\mathcal{F}$ associated with X are replaced by $X = x$. Throughout the paper, we use a specific cluster causal diagram $\mathcal{G}_{\text{SFM}}$ known as the standard fairness model (SFM) (Plečko and Bareinboim, 2024) over endogenous variables $\{X, Z, W, Y\}$, shown in Fig. 2a, representing the protected attribute, confounders, mediators, and the outcome, respectively.

Our key goal is to decompose the average difference in mortality rates between majority and minority groups, as given in Eq. 1-2. The average difference in mortality, written $\mathbb{E}[Y \mid X = x_1] - \mathbb{E}[Y \mid X = x_0]$, where Y is the outcome, x_1 the majority group, x_0 minority, can be decomposed as (Zhang and Bareinboim, 2018; Plečko and Bareinboim, 2024):

$$\begin{aligned}\mathbb{E}[Y \mid X = x_1] - \mathbb{E}[Y \mid X = x_0] = \mathbb{E}[Y_{x_1, W_{x_0}} - Y_{x_0} \mid X = x_0] \quad &\text{(direct)} \\ - \mathbb{E}[Y_{x_1, W_{x_0}} - Y_{x_1} \mid X = x_0] \quad &\text{(indirect)} \\ - \mathbb{E}[Y_{x_1} \mid X = x_0] - \mathbb{E}[Y_{x_1} \mid X = x_1]. \quad &\text{(confounded)}\end{aligned} \tag{6}$$

where we label the terms on the RHS $x\text{-DE}_{x_0,x_1}(y \mid x_0)$, $x\text{-IE}_{x_1,x_0}(y \mid x_0)$, and $x\text{-CE}_{x_1,x_0}(y)$, respectively. The difference between $\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0} \mid X = x_0]$ captures the effect of changing X from minority group x_0 to majority group x_1 along the direct causal pathway while keeping the mediators at their natural level W_{x_0}, averaged across all minority patients (represented by the conditioning $X = x_0$ in the expectation). The difference $\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_1} \mid X = x_0]$ captures the effect of changing X from majority group x_1 to minority group x_0 along the indirect causal pathway while keeping $X = x_1$ along the direct

causal path, average across individuals. For the indirect effect, we considered *the reverse transition* of changing $x_1 \to x_0$ (as opposed to $x_0 \to x_1$ for the direct effect), and this difference was subtracted from the direct effect quantity. Finally, the confounded term $\mathbb{E}[Y_{x_1} \mid X = x_0] - \mathbb{E}[Y_{x_1} \mid X = x_1]$ compares the effect of setting $X = x_1$ along direct and indirect pathways for the minority group x_0 vs. the majority group x_1. The marginal difference $\mathbb{E}[Y \mid X = x_1] - \mathbb{E}[Y \mid X = x_0]$ is obtained by subtracting the above described quantifications of indirect and confounded effects from the quantification of the direct effect as shown in Eq. 6. In Eqs. 3-4 we report the $x\text{-DE}_{x_0,x_1}(y \mid x_0)$ effect, and the negative of indirect, and confounded effects, $-x\text{-IE}_{x_1,x_0}(y \mid x_0)$ and $-x\text{-CE}_{x_1,x_0}(y)$, respectively. This way of reporting the results allows us to represent the TV measure in a simplified way, written as

$$\text{TV} = \underbrace{x\text{-DE}_{x_0,x_1}(y \mid x_0)}_{\text{direct}} + \underbrace{(-x\text{-IE}_{x_1,x_0}(y \mid x_0))}_{\text{indirect}} + \underbrace{(-x\text{-CE}_{x_1,x_0}(y))}_{\text{confounded}}. \tag{7}$$

For estimating the effects from finite sample, we derived the efficient influence functions for each term appearing in Eq. 7, and performed one-step debiasing (Kennedy, 2024) to obtain asymptotically normal estimators, for which the uncertainty estimation can be obtained using the normal approximation. Sensitivity analysis for the impact of data missingness on the inference of results is described in Appendix E, while in Appendix C we perform a sensitivity analysis to investigate the effect of overlap on inference.

3.3 Interaction Testing and Heterogeneous Effects

When decomposing the marginal disparity into direct, indirect, and confounded effects, a transition $x_0 \to x_1$ along the direct effect needs to be considered, and a reverse transition $x_1 \to x_0$ along the indirect effect is subtracted from it. The reason why a reverse transition needs to be considered is the possible existence of interactions, as noted by (VanderWeele, 2015). Following (Plecko, 2024), an absence of an interaction would imply that the structural mechanism f_y of Y can be written as:

$$f_y(x, z, w, u_y) = f_y^{(1)}(x, u_y) + f_y^{(2)}(w, u_y) + f_y^{(3)}(z, u_y), \tag{8}$$

where $f_y^{(1)}, f_y^{(2)}$, and $f_y^{(3)}$ are the structural functions corresponding to effects of covariates X, W, and Z, respectively. For binary outcomes, the absence of interactions is considered on the risk-scale, meaning that the structural mechanism f_y which returns a binary value is replaced by $p_y(x, z, w) = P(f_y(x, z, w, U_y) = 1)$. For the binary case, the absence of any interactions would be written as this would be given by the condition:

$$p_y(x, z, w) := P(f_y(x, z, w, U_y) = 1) = \exp\{p_y^{(1)}(x) + p_y^{(2)}(w) + p_y^{(3)}(z)\}, \tag{9}$$

where $p^{(i)}$ are different functions. In case of no interactions (Eq. 9), we have the following implications:

$$x\text{-DE}_{x_0,x_1}(y \mid x_0) = x\text{-DE}_{x_1,x_0}(y \mid x_0) \tag{10}$$
$$x\text{-IE}_{x_0,x_1}(y \mid x_0) = x\text{-IE}_{x_1,x_0}(y \mid x_0) \tag{11}$$
$$x\text{-CE}_{x_0,x_1}(y) = x\text{-CE}_{x_1,x_0}(y) \tag{12}$$

We follow the approach of (Plecko, 2024) as described above and perform non-parametric interaction testing by performing a hypothesis test for each of the Eqs. 10-12. The obtained p-values for each interaction are provided in Appendix B. For quantifying heterogeneous effects, we considered four different age groups $C_1, \ldots, C_4$, corresponding to 18-54, 55-64, 65-74, and 74-100 years of age. These groups were chosen as approximate quartiles of age in the data. For admission categories, we considered three admission types, namely medical admissions d_{med}, emergency surgery admissions d_{ems}, and elective surgery admissions d_{els}. For investigating heterogeneity of direct effects, we computed the quantity

$$\text{DE}_{x_0,x_1}(y \mid C_i, d, x_0) = \mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0,W_{x_0}} \mid \text{age} \in C_i,\ \text{admission} = d, X = x_0] \tag{13}$$

for different values of C_i, d (reported in Fig. 4a). For heterogeneity of indirect effects, we investigated the quantity

$$-\text{IE}_{x_1,x_0}(y \mid \text{age} \in C_i, x_0) = \mathbb{E}[Y_{x_1,W_{x_1}} - Y_{x_1,W_{x_0}} \mid \text{age} \in C_i, X = x_0] \tag{14}$$

for different values of C_i (see Appendix D). The effects $\text{DE}_{x_0,x_1}(y \mid C_i, d, x_0)$ and $\text{IE}_{x_1,x_0}(y \mid C_i, x_0)$ were estimated using causal forests (Wager and Athey, 2018), while the uncertainty estimates were obtained using bootstrap.

3.4 Tipping Over Hypothesis and Population Risks

The Australian Bureau of Statistics provides census-based values of population in each age group for years 2016, 2021 (Australian Bureau of Statistics, 2021b). Based on this data, we performed piecewise linear interpolation in each age group for both the overall and majority populations to determine the population counts between 2018 and 2024. After performing the imputation, we were able to compute the number of minority and majority persons who are possibly at risk in each year t and age group a, denoted as $N(x_0, a, t), N(x_1, a, t)$, respectively. The number of minority and majority patients admitted in year t and age group a to the ICU for diagnosis group d was labeled $n(x_0, a, d, t), n(x_1, a, d, t)$, respectively. The indicator of whether a person is admitted is labeled with I. The probability (or risk) of admission for a diagnosis $D = d$, in age group a, demographic group x, and year t, denoted by $P(I = 1 \wedge D = d \mid \text{age} = a, \text{year} = t, X = x)$, is given by:

$$P(I = 1 \wedge D = d \mid \text{age} = a, \text{year} = t, X = x) = \frac{n(x, a, d, t)}{N(x, a, t)}. \tag{15}$$

For any conditioning event E, we have that E-specific risk to be defined as

$$P(I = 1 \wedge D = d \mid E, do(X = x)) = \sum_{a,t} P(I = 1 \wedge D = d \mid \text{age} = a, \text{year} = t, X = x) P(a, t \mid E) \tag{16}$$

where $P(a, t \mid E)$ is defined as

$$P(a, t \mid E) = \frac{\sum_{a,t \in E} N(x_0, a, t) + N(x_1, a, t)}{\sum_{a,t} N(x_0, a, t) + N(x_1, a, t)}. \tag{17}$$

Then, to obtain the effect of minority status on the risk of ICU admission for any conditioning event E, we can compute the E-specific risk ratio:

$$\text{RR}(d|E) = \frac{P(I = 1 \wedge D = d \mid E, do(X = x_0))}{P(I = 1 \wedge D = d \mid E, do(X = x_1))}. \tag{18}$$

In Fig. 5, we reported the logarithms of $\text{RR}(d \mid \text{year} = t)$ for year t varying in $2018, \ldots, 2024$, and for different diagnoses d. In Fig. 6a we reported $\text{RR}(d \mid \text{age} \in C_i)$ for medical diagnoses d_{med}, emergency surgery diagnoses d_{ems}, elective surgery diagnoses d_{els}, and across age groups C_i (18-49, 50-64, 65-74, 75-100 years old).

4 Discussion

In this paper, we introduced a systematic approach for analyzing health disparities, based on the tools of causal fairness analysis (Plečko and Bareinboim, 2024). The framework was illustrated through an analysis of racial/ethnic disparities in ICU outcomes between minority and majority groups in Australia and the United States (see Fig. 1). Our investigation demonstrated that commonly used, statistical measures to quantify disparity are insufficient for investigating health equity. In the Australian cohort, the minority group had a higher average mortality rate than the majority group, while this was reversed in the US cohort. However, when taking a more fine-grained, causal perspective on data analysis, the studied effects (direct, indirect, spurious) had the same signs and were consistent across populations.

Using a causal lens also allowed us to investigate different possible pathways that may transmit change between minority status and the outcome of interest. When considering confounded effects, we found that minority race/ethnicity was associated with lower age at admission in both countries (Martin et al, 2003), which in turn reduced the risk of mortality. This effect is most likely driven by increased chronic health problems, and/or lack of access to primary care for the minority groups. For the indirect effects, we found that minority groups had worse chronic health, were more likely to be admitted for urgent (non-elective) and medical reasons, and had higher illness severity levels (after removing the age effect). All of these factors are known to increase the mortality risk. At the same time, surprisingly, we

found that along the direct effect, minority patients showed better survival, when all other variables were held constant (*ceteris paribus*). A further analysis of effect interactions showed that this direct effect was heterogeneous across admission types. For medical admissions, there was a pronounced protective direct effect for the minority group, while this was not the case for surgical admissions. We hypothesized that the protective direct effect of race/ethnicity, which applied to medical admissions, may be due to the fact that minority patients have poor access to primary healthcare, as reported in previous literature (Davy et al, 2016). The causal mechanism for such a protective effect would be that worse access to primary care results in a higher prevalence of ICU admission for a specific diagnosis, and that such increased prevalence reduces the overall risk of death in that group, since increased prevalence also implies that less severe cases reach the ICU. A related phenomenon of increased prevalence and reduced mortality has been observed in the literature on sex-related ICU disparities (Modra et al, 2022). In this context, we analyzed the baseline risk of ICU admission in Australia. We showed that the increase in risk for the minority group was highest in the group of medical admissions (compared to the baseline risk for the majority group), and significantly higher than for surgical admissions. Such increase in risk constitutes a selection bias (or left-censoring of majority group patients), which may explain the observed direct effect. Accordingly, we demonstrated that the pattern of increase in the risk of ICU admission was statistically related to the strength of the protective direct effect observed for different diagnostic and age groups. In a separate analysis of readmissions, we found that minority patients were in fact more likely to be readmitted (Soto et al, 2013), and this effect was also strongest within the group of medical admissions. The fact that Indigenous patients showed increased risk of admission, improved survival, and also a higher chance of readmission, supports the interpretation that under-utilization of primary health care may be one of the causes of our findings. This motivated us to construct the Indigenous Intensive Care Equity (IICE) Radar, which monitors the increase in risk of ICU admission for Indigenous patients across different geographical areas. The increase in risk of ICU is a hypothesized proxy of lack of access and reduced utilization of primary care. The construction of the Radar opens the door to important future studies with possibly significant public health implications.

An important finding of our study is that, when considering direct, indirect, and spurious effects, racial and ethnic disparities in ICU outcome are a consequence of differences that happen prior to the time the patient enters the ICU, with key factors being (i) worse chronic health and lower age at admission; (ii) higher risk of a non-elective, urgent admission; (iii) worse access to primary care for earlier treatment of preventable ICU admissions.

Another important takeaway of our study is that the "sign" of a causal effect estimate is not a definitive way to conclude whether a protected group is discriminated against. For instance, the spurious association of minority race/ethnicity with age, which results in better outcome, showed a protective effect on outcome in our data. Similarly, for the direct effect, we also observed a protective effect of minority race/ethnicity. However, upon interpreting these effects, we noted that they are most likely a consequence of socioeconomic disadvantage and worse access to healthcare. Therefore, critical care needs to be interpreted within a multi-layered system of healthcare, and the filtering and selection bias of populations that require critical care has direct implications on the health disparities observed in ICU outcomes.

A major strength of our work is the size and the heterogeneity of the data we studied. We studied almost 1.1 million ICU admissions across two countries on different continents, and our findings were robust and consistent across these countries. A second strength is the use of a systematic framework for analyzing health inequities, which allows us to both decompose the observed disparity in average mortality rates into its constitutive causal elements, and also allows us to investigate whether there are significant interactions between different causal pathways. Thirdly, we were able to establish hypotheses that explained why differences along direct, indirect, and spurious pathways occur, and we tested a tipping-over hypothesis to explain the protective direct effect for minority patients. We further connected the quantification of the direct effects with external data on the baseline risk of ICU admission, adding further evidence to support our explanation of the direct effects, and constructed a monitoring tool that may provide the basis for important public health policy in the future.

In terms of limitations, we acknowledge the observational nature of our study, and note that some relevant confounders such as socioeconomic status (missing in US data), or chronic health status (missing in Australian data) were partially included in our analyses. Still, we were able to demonstrate that tools of causal inference may help uncover important patterns in health equity investigations. Furthermore, as elaborated in the main text, the missing confounders would likely not explain away the effects we studied, but rather make the protective direct effect for minority patients even more pronounced. Secondly, we note that our analysis of baseline risk of ICU admission did not include the overall prevalence of different diseases in the population but focused on rates of admission to ICU

according to different diagnoses. However, our findings on the increased admission, improved survival, and increased readmission do seem to support the hypothesis that the observed differences are related to under-utilization of primary care.

5 Conclusion

In a large study of almost 1.1 million ICU admissions in Australia and the United States, we used the framework of Causal Fairness Analysis to investigate whether tools of causal reasoning can help to disentangle the mechanisms underlying disparities in mortality according to race and ethnicity among critically ill patients. We found that three different causal effects explained the disparity: (i) minority patients were admitted younger on average, decreasing the mortality rate; (ii) minority patients had worse chronic health and were more likely to be admitted for non-elective and medical reasons, increasing the mortality risk; (iii) there was a protective direct effect for minority group patients admitted for medical reasons, indicating a decrease in mortality compared to the majority group, when all other variables were kept equal. The last, protective direct effect is most likely explained through a causal mechanism in which minority patients are, due to worse access to primary health care, more likely to end up in ICU for less severe and preventable conditions, which then spuriously reduces the mortality risk. Finally, we emphasize that the novel framework used in this manuscript can be applied across a range of problems in health equity, as it provides a systematic way of causally explaining health disparities, while being compatible with the modern tools of large-scale data analysis and causal inference.

Acknowledgements

Declarations

Indigenous Data Statement. Indigenous data is that which is "generated, intentionally or not, by, about, or for Aboriginal and Torres Strait Islander people". It also refers to "Information, in any format or medium, collected, analysed, stored, and interpreted within the context of Indigenous individuals, collectives, populations, entities, lifeways, cultures, knowledge systems, lands, biodiversity, water and other resource" (IDN, 2024). Given the history of exploitation of Indigenous people in research and data collection practices in Australia, it is therefore critical that Indigenous knowledges and approaches are prioritized when seeking or using data for research from Indigenous populations (Kukutai and Taylor, 2016; Rainie et al, 2019; Walker and Davies, 2019).

The datasets in this study that were obtained from ANZICS and contain information relating to Aboriginal and Torres Strait Islander people in Australia. Data has been used considering the international FAIR principles of; Findable, Accessible, Interoperable and Reusable, and CARE principles of; Collective benefit, Authority to control, Responsibility and Ethics (Alliance, 2019). These principles will be embedded to ensure that Indigenous data governance and Indigenous knowledges are prioritized, but also ensure that the results benefit Indigenous peoples.

- **Funding:** Author LAC is funded by the National Institute of Health through R01 EB017205, DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, and the National Science Foundation through ITEST #2148451. Other authors were funded by their respective institutions.

- **Conflict of interest/Competing interests:** Authors declare no competing interests or conflicts of interest.
- **Ethics approval and consent to participate:** The research in this paper was approved by the ethics committee of the Alfred Hospital, Melbourne (Project #661/24). This study utilized retrospective, de-identified data, and therefore, informed consent was waived by the ethics committee in accordance with applicable regulations and guidelines.
- **Consent for publication:** All authors consent to the publication of the manuscript.
- **Data availability:** The MIMIC-IV dataset is publicly available from Physionet (Goldberger et al, 2000). The ANZICS APD is available through the Australian and New Zealand Intensive Care Society (ANZICS) upon request.
- Materials availability: N/A.
- **Code availability:** All the code used in the paper is available in the Github repository https://github.com/dplecko/RaceMortalityICU. The README file provides instructions for installation and reproducing the results, and each script contains comments explaining the code logic.
- **Author contributions:** DP, EB, LAC, and RB developed the study design. DP carried out the statistical analyses and data processing. DP wrote the manuscript. PS, LAC, and DPilcher helped write the manuscript and interpret the results. AF, SA, and DD provided discussions on the interpretation of the results and their implications for minority health.

References

Alliance GID (2019) Care principles for indigenous data governance.[pdf] research data alliance. international indigenous data sovereignty interest group

Australian Bureau of Statistics (2021a) Estimates of aboriginal and torres strait islander australians. URL https://www.abs.gov.au/statistics/people/aboriginal-and-torres-strait-islander-peoples, accessed: 2024-07-08

Australian Bureau of Statistics (2021b) National, state and territory population - june 2021. URL https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/jun-2021, accessed: 2024-07-08

Australian Bureau of Statistics (2021c) Socio-economic indexes for areas (seifa), australia. URL https://www.abs.gov.au/statistics/people/people-and-communities/socio-economic-indexes-areas-seifa-australia/latest-release, aBS Website, accessed 6 December 2024

Bareinboim E, Pearl J (2016) Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences 113(27):7345–7352

Bareinboim E, Correa JD, Ibeling D, et al (2022) On pearl's hierarchy and the foundations of causal inference. In: Probabilistic and Causal Inference: The Works of Judea Pearl, 1st edn. Association for Computing Machinery, New York, NY, USA, p 507–556

Bennett N, Plečko D, Ukor IF, et al (2023) ricu: R's interface to intensive care data. GigaScience 12:giad041
Braveman P (2014) What are health disparities and health equity? we need to be clear. Public health reports 129(1_suppl2):5–8
Braveman P, Arkin E, Orleans T, et al (2017) What is health equity? and what difference does a definition make? robert wood johnson foundation
Braveman PA, Kumanyika S, Fielding J, et al (2011) Health disparities and health equity: the issue is justice. American journal of public health 101(S1):S149–S155
Centers for Disease Control and Prevention (CDC) (2024) Health equity fact sheet. URL https://www.cdc.gov/healthequity/core/fact-sheet/index.html
Charlson ME, Pompei P, Ales KL, et al (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases 40(5):373–383
Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794
Commonwealth Fund (2023) Global perspective on u.s. health care. URL https://www.commonwealthfund.org/publications/issue-briefs/2023/jan/us-health-care-global-perspective-2022
Crump RK, Hotz VJ, Imbens GW, et al (2009) Dealing with limited overlap in estimation of average treatment effects. Biometrika 96(1):187–199
Davy C, Harfield S, McArthur A, et al (2016) Access to primary health care services for indigenous peoples: a framework synthesis. International journal for equity in health 15:1–9
Dwyer-Lindgren L, Kendrick P, Kelly YO, et al (2022) Life expectancy by county, race, and ethnicity in the usa, 2000–19: a systematic analysis of health disparities. The Lancet 400(10345):25–38. https://doi.org/10.1016/s0140-6736(22)00876-5, URL http://dx.doi.org/10.1016/S0140-6736(22)00876-5
European Union (EU) (2024) Strong health systems, universal health coverage and social participation. URL https://belgian-presidency.consilium.europa.eu/en/events/towards-health-equity-strong-health-systems-universal-health-coverage-and-social-participation/
Evans RS (2016) Electronic health records: Then, now, and in the future. Yearbook of Medical Informatics 25(S 01):48–61
Goldberger AL, Amaral LA, Glass L, et al (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation 101(23):e215–e220
Health TLP (2023) Health and inequity in australia. The Lancet Public Health 8(8):e575. https://doi.org/10.1016/s2468-2667(23)00157-3, URL http://dx.doi.org/10.1016/S2468-2667(23)00157-3
Healthy People 2030 (2023) Healthy people 2030 questions & answers. URL https://health.gov/our-work/national-health-initiatives/healthy-people/healthy-people-2030/questions-answers
Indigenous Data Network (2024) Indigenous Data Network of Australia. URL https://idnau.org/, accessed: 2024-11-15
Johnson A, Bulgarelli L, Pollard T, et al (2020) Mimic-iv. PhysioNet Available online at: https://physionet org/content/mimiciv/10/(accessed August 23, 2021) pp 49–55
Kennedy EH (2024) Semiparametric doubly robust targeted double machine learning: a review. Handbook of Statistical Methods for Precision Medicine pp 207–236
Knaus WA, Wagner DP, Draper EA, et al (1991) The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. Chest 100(6):1619–1636
Kukutai T, Taylor J (2016) Indigenous data sovereignty: Toward an agenda. ANU press
Lei L, D'Amour A, Ding P, et al (2021) Distribution-free assessment of population overlap in observational studies. Unpublished working paper
Malaspinas AS, Westaway MC, Muller C, et al (2016) A genomic history of aboriginal australia. Nature 538(7624):207–214
Martin GS, Mannino DM, Eaton S, et al (2003) The epidemiology of sepsis in the united states from 1979 through 2000. New England Journal of Medicine 348(16):1546–1554
McGowan SK, Sarigiannis KA, Fox SC, et al (2022) Racial disparities in icu outcomes: a systematic review. Critical care medicine 50(1):1–20
Modra LJ, Higgins AM, Pilcher DV, et al (2022) Sex differences in mortality of icu patients according to diagnosis-related sex balance. American journal of respiratory and critical care medicine 206(11):1353–1360
Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, 2nd edition, 2009
Plecko D (2024) Interaction testing in variation analysis. arXiv preprint arXiv:241108861
Plečko D, Bareinboim E (2024) Causal fairness analysis: a causal toolkit for fair machine learning. Foundations and Trends® in Machine Learning 17(3):304–589

Plecko D, Bareinboim E (2024) Mind the gap: A causal perspective on bias amplification in prediction & decision-making. arXiv preprint arXiv:240515446

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Rainie SC, Kukutai T, Walter M, et al (2019) Indigenous data sovereignty. African Minds and the International Development Research Centre (IDRC)

Secombe P, Millar J, Litton E, et al (2023) Thirty years of anzics core: A clinical quality success story. Critical Care and Resuscitation 25(1):43–46

Soto GJ, Martin GS, Gong MN (2013) Healthcare disparities in critical illness. Critical care medicine 41(12):2784–2793

Tishkoff SA, Reed FA, Friedlaender FR, et al (2009) The genetic structure and history of africans and african americans. science 324(5930):1035–1044

U.S. Department of Health & Human Services (HHS) (2024a) Advancing equity at hhs. URL https://www.hhs.gov/equity/index.html

U.S. Department of Health & Human Services (HHS) (2024b) Fact sheet: Advancing health equity across hhs. URL https://www.hhs.gov/equity/fact-sheet-advancing-health-equity-across-hhs/index.html

VanderWeele T (2015) Explanation in causal inference: methods for mediation and interaction. Oxford University Press

Vincent JL, Moreno R, Takala J, et al (1996) The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix)

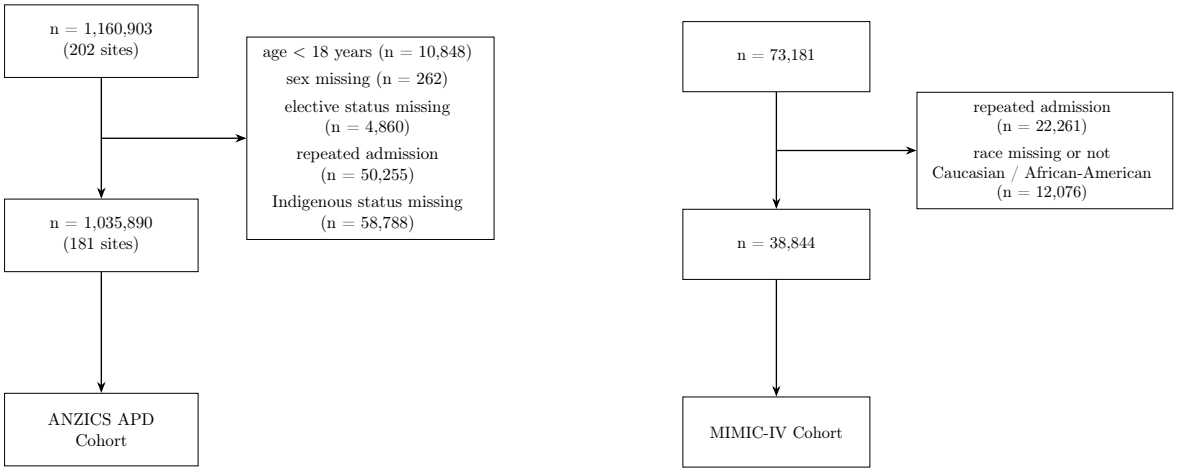
Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523):1228–1242

Walker B, Davies T (2019) The State of Open Data: Histories and Horizons. African Minds

World Health Organization (WHO) (2024) Who releases new guidance on monitoring the social determinants of health equity. URL https://www.who.int/news/item/19-02-2024-who-releases-new-guidance-on-monitoring-the-social-determinants-of-health-equity

Zhang J, Bareinboim E (2018) Fairness in decision-making—the causal explanation formula. In: Proceedings of the AAAI Conference on Artificial Intelligence

Appendix A Patient Information & Filtering



(a) ANZICS APD filtering.

(b) MIMIC-IV filtering.

Fig. A1: Study flowchart of patient filtering steps for (a) ANZICS APD and (b) MIMIC-IV.

Variable	Reported	ANZICS APD (Majority)	ANZICS APD (First Nations)	p-value
Cohort size	n	997154	38736	
Age (years)	Median (IQR)	66.58 (52.88-76.08)	50.6 (37.09-61.6)	< 0.001
Admission type				< 0.001
- Medical	%	45	69	
- Surgical	%	55	31	
Mortality	n (%)	74820 (7.5%)	3064 (7.9%)	0.003
ICU LOS (days)	Median (IQR)	1.76 (0.92-3.25)	2.02 (1.01-3.96)	< 0.001
Hospital LOS (days)	Median (IQR)	7.71 (4.05-14.06)	7.16 (3.48-14.2)	< 0.001
Sex (Male)	%	56	51	< 0.001
Ventilated	n (%)	306332 (30.7%)	15164 (39.1%)	< 0.001
APACHE-III Score	Mean, Med. (IQR)	50.7, 47 (35-62)	51.6, 47 (33-66)	0.063
APACHE-III Risk of Death	Mean, Med. (IQR)	0.12, 0.05 (0.02-0.14)	0.14, 0.06 (0.02-0.16)	< 0.001

Table A1: Comparison of patient characteristics on ANZICS APD.

Variable	Reported	MIMIC-IV (White)	MIMIC-IV (African-American)	p-value
Cohort size	n	34204	4640	
Age (years)	Median (IQR)	67 (55-78)	60 (47-72)	< 0.001
Admission type				< 0.001
- Medical	%	60	75	
- Surgical	%	40	25	
Mortality	n (%)	2627 (7.7%)	318 (6.9%)	< 0.001
ICU LOS (days)	Median (IQR)	1.85 (1.08-3.37)	1.8 (1-3.26)	< 0.001
Hospital LOS (days)	Median (IQR)	6.5 (3.91-10.9)	6.41 (3.68-11.54)	0.460
Sex (Male)	%	56	46	< 0.001
Ventilated	n (%)	13692 (40.0%)	1480 (31.9%)	< 0.001
SOFA				< 0.001
- Respiratory	Median (IQR)	2 (1-3)	2 (1-3)	
- Coagulation	Median (IQR)	0 (0-1)	0 (0-1)	
- Hepatic	Median (IQR)	0 (0-1)	0 (0-1)	
- Cardio	Median (IQR)	1 (1-1)	1 (0-1)	
- CNS	Median (IQR)	0 (0-1)	0 (0-1)	
- Renal	Median (IQR)	0 (0-1)	0 (0-1)	
- Total	Median (IQR)	4 (2-6)	4 (2-6)	

Table A2: Comparison of patient characteristics on MIMIC-IV.

Appendix B Interaction Testing

Interaction Test	ANZICS APD (AU)	MIMIC-IV (US)
TE $\otimes$ SE	$< 0.01^*$	0.39
DE $\otimes$ IE	$< 0.01^*$	0.19
DE $\otimes$ SE	0.09	0.35
IE $\otimes$ SE	$< 0.01^*$	0.02^*
DE $\otimes$ IE $\otimes$ SE	0.57	0.71

Table B3: Interaction Testing for ANZICS APD and MIMIC-IV datasets.

Appendix C Investigating Overlap

An important assumption required for correct causal effects estimation is known as overlap:

$$\delta < P(X = 1 \mid Z = z, W = w) < 1 - \delta, \tag{C1}$$

for some $\delta > 0$. In this appendix, we investigate the overlap assumption in the datasets we analyzed. We abbreviate the quantity $P(X = 1 \mid Z = z, W = w)$ by $e_{x_1}(z, w)$, and similarly $P(X = 0 \mid Z = z, W = w)$ by $e_{x_0}(z, w)$. The minimum of the two propensity weights $e_{\min}(z, w)$ is defined as

$$e_{\min}(z, w) := \min\{e_{x_0}(z, w), e_{x_1}(z, w)\}. \tag{C2}$$

The notation $e_x(Z, W)$ denotes a random variable depending on random values Z, W, as opposed to a fixed value $e_x(z, w)$ for $Z = z, W = w$. To investigate the validity of the overlap assumption in Eq. C1, we first estimate the propensity weights $\hat{e}_x(Z_i, W_i)$ for each of our data samples (Z_i, W_i). Then, we perform a sensitivity analysis using quantile based trimming of the propensity weights (Crump et al, 2009). Let $Q(q; e_{\min}(Z, W))$ denote the lower q quantile of the distribution of the propensity weights $e_{\min}(Z, W)$. For each dataset, and each quantile $q \in \{1\%, \ldots, 5\%\}$, we select all the samples with values above the lower q quantile, and construct the dataset

$$\mathcal{D}(q) = \{(X_i, Z_i, W_i, Y_i) \;\; \forall i : \hat{e}_{\min}(Z_i, W_i) > Q(q; \hat{e}_{\min}(Z, W))\}. \tag{C3}$$

We then perform the decomposition of the TV measure as in Eq. 6 for the subset of the data $\mathcal{D}(q)$ with the q-quantile of samples with extreme propensity weights removed. For the dataset $\mathcal{D}(q)$, for which the overlap criterion in Eq. C1 should hold with the value of $\delta = Q(q; \hat{e}_{\min}(Z, W))$, we compute the 95% O-value using the difference-in-tails method (DiT) (Lei et al, 2021), labeled $\hat{\mathcal{O}}(q)$, which provides an upper bound on the overlap value δ for the dataset $\mathcal{D}(q)$, in a distribution-free way. To test for overlap violations, we use the hypothesis test that compares if $\hat{\mathcal{O}}(q)$ is smaller than the nominal overlap bound $Q(q; \hat{e}_{\min}(Z, W))$ of the trimmed dataset $\mathcal{D}(q)$. If so, the test detects overlap violations and signals that a stronger trimming procedure is necessary.

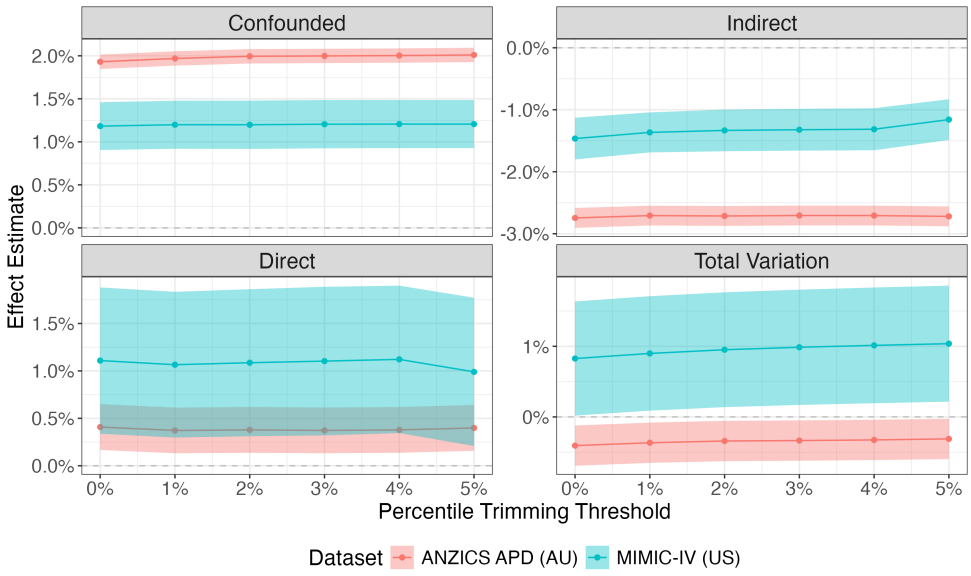


Fig. C2: Sensitivity analysis of the impact of quantile-based trimming of propensity weights on TV decompositions on ANZICS APD and MIMIC-IV.

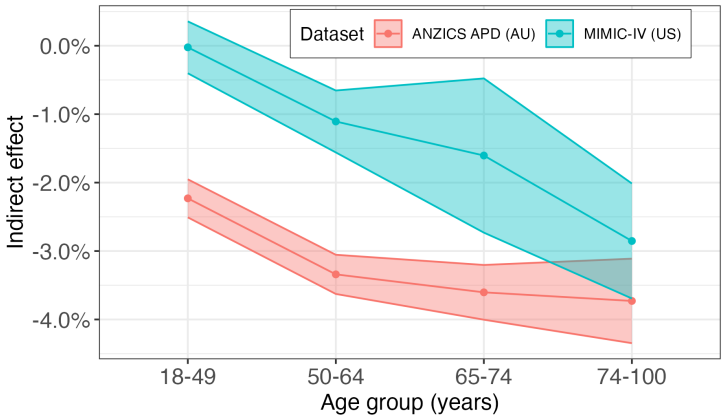
Fig. C2 provides the results of the sensitivity analysis, and demonstrates that the causal decomposition is not significantly affected by the trimming of propensity weights over different thresholds q. Furthermore, for both datasets and at each threshold q, we found that the O-values were greater than the nominal overlap condition after trimming, namely $\hat{\mathcal{O}}(q) > Q(q; \hat{e}_{\min}(Z, W))$, meaning that no significant overlap violations were detected (trimming was strong enough). Putting everything together, we conclude that there is no evidence that the TV decompositions are affected by overlap violations.

Appendix D Indirect Effect Heterogeneity

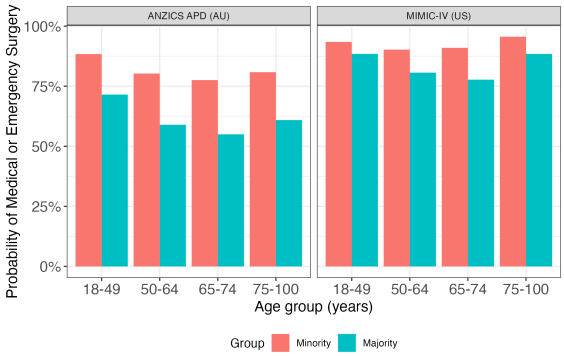
The interaction testing analysis performed in Sec. 2.4 demonstrated that important interactions exist between direct and indirect pathways, and also indirect and spurious pathways. While the heterogeneity of direct effects was investigated in the main text, in this appendix we focus on the heterogeneity of indirect effects according to age, as a consequence of significant interactions of spurious and indirect paths. For the four age groups 18-54, 55-64, 65-74, and 74-100, labeled C_1 to C_4, and corresponding approximately to the quartiles of age in the data, we compute the indirect effects

$$-\text{IE}_{x_1,x_0}(y \mid \text{age} \in C_i, x_0) = \mathbb{E}[Y_{x_1,W_{x_1}} - Y_{x_1,W_{x_0}} \mid \text{age} \in C_i, X = x_0]. \tag{D4}$$

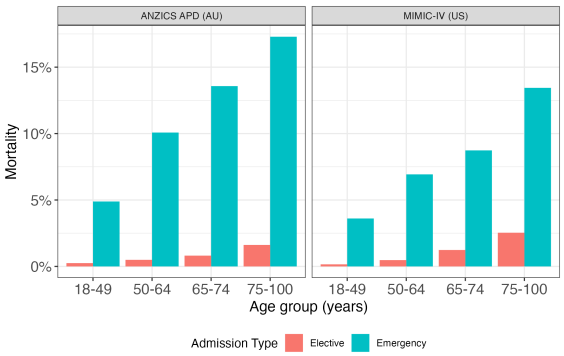
The effects are estimated for both the Australian and US data, and visualized in Fig. D3a. In both datasets, the strength of the indirect increases with age, and the risk of death for majority group patients is reduced along the indirect path. After establishing these effects, we need to try to understand



(a) Indirect effect heterogeneity across age groups on ANZICS APD and MIMIC-IV.



(b) Effect of minority status on admission type.



(c) Effect of admission type on mortality.

Fig. D3: Understanding the heterogeneity of indirect effects on ANZICS APD and MIMIC-IV data.

mechanistically what drives this heterogeneity. In Fig. D3b we plot how the proportion of medical and emergency surgery admissions (these admissions are jointly referred to as urgent) changes across age groups C_i and minority status. In Fig. D3c, we plot how the mortality rate for urgent and elective admissions changes across age groups C_i. Fig. D3b illustrates that in every age group C_i, minority patients are more likely to be admitted for an urgent condition. Fig. D3c illustrates that while the risk of death increases with age for both urgent and elective admissions, the difference in risk of urgent vs. elective admissions becomes larger with age. These observations explain why the indirect effect, along which minority patients are more likely to die, is more pronounced with older age.

Appendix E Missing Value Sensitivity Analysis

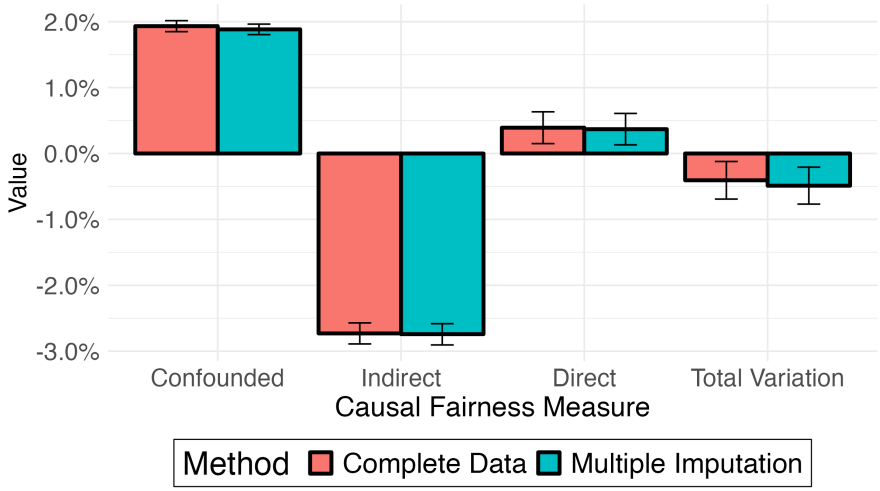


Fig. E4: Sensitivity analysis of how the imputation of missing values of X affects the TV decomposition on the ANZICS APD data.

In this appendix, we investigate the effect of missing values on the data analysis. In the ANZICS APD, of the overall cohort under investigation, $n = xyz$ patients had no reported Indigenous status. To investigate the effect of this data missingness on the decompositions of the TV measure, we proceed as follows. Let R denote the missingness indicator. We assume that the data is missing at random (MAR), meaning that the missingness pattern depends on the other observed variables Z, W, Y, and not on the value of X itself, i.e.,

$$R \perp\!\!\!\perp X \mid Z, W, Y. \tag{E5}$$

Written differently, this conditioning can also be given as

$$P(R = 1 \mid X = x_1, Z = z, W = w, Y = y) = P(R = 1 \mid X = x_0, Z = z, W = w, Y = y). \tag{E6}$$

For a fixed value of $Z = z, W = w, Y = y$, the missingness of X does not depend on the actual value of X. A standard consequence of the MAR assumption is the fact that

$$P(X = x \mid R = 1, Z = z, W = w, Y = y) = P(X = x \mid R = 0, Z = z, W = w, Y = y). \tag{E7}$$

In words, the distribution of X given Z, W, Y is the same within the subsets of data where X is unobserved ($R = 1$) and where X is observed ($R = 0$). Therefore, we perform the following sensitivity analysis. We first regress X on Z, W, Y using xgboost (Chen and Guestrin, 2016) on the complete subset of data with X recorded. Then, for the subset where X is missing, we impute the missing values of X based on the observed values $Z = z, W = w, Y = y$. We repeat the imputation process ten times, obtaining imputed datasets $\mathcal{D}_1, \ldots, \mathcal{D}_{10}$. We repeat the TV decomposition on each dataset $\mathcal{D}_i$, and compute the effect estimates and their confidence interval. These results (based on imputed data) are then compared to the data analysis focusing on complete data only, to understand how much additional uncertainty is added once missingness is taken into account. The results of this experiment are shown in Fig. E4, and indicate that the data missingness has a negligible impact on the results.