Causal Analysis of the TOPCAT Trial: Spironolactone for Preserved Cardiac Function Heart Failure

Francesca E. D. Raimondi
causaLens

Tadhg O'Keeffe
causaLens

Hana Chockler
causaLens
King's College London

Andrew R. Lawrence
causaLens

Tamara Stemberga
causaLens

Andre Franca
causaLens

Maksim Sipos
causaLens

Javed Butler
Baylor Scott & White Research Institute, Dallas, TX
University of Mississippi, Jackson, MS

Shlomo Ben-Haim
Hobart Group

Abstract

We describe the results of applying causal discovery methods on the data from a multi-site clinical trial, on the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist (TOPCAT). The trial was inconclusive, with no clear benefits consistently shown for the whole cohort. However, there were questions regarding the reliability of the diagnosis and treatment protocol for a geographic subgroup of the cohort. With the inclusion of medical context in the form of domain knowledge, causal discovery is used to demonstrate regional discrepancies and to frame the regional transportability of the results. Furthermore, we show that, globally and especially for some subgroups, the treatment has significant causal effects, thus offering a more refined view of the trial results.

1 Introduction

Complex trials with heterogeneities in the data pose a significant challenge for traditional techniques of statistical analysis. Suboptimal patient cohort selection and recruitment, paired with ineffective monitoring of patients, are among the main causes for high trial failure rates: only 10% of the compounds entering a clinical trial reaches the market [1]. Phase 2 clinical studies, albeit a poor predictor of drug success ($> 30\%$ and $> 58\%$ trial failure rates in phase 2 and 3, respectively) are critical in determining drug costs [2]. Results of multi-site clinical trials often show significant differences between the trial sites. There can be a variety of reasons for this, including different treatment benefits among different types of participants or site anomalies that could invalidate the study, such as suspected failures to assign the correct diagnosis and to adhere to treatment protocols.

The TOPCAT clinical trial is a randomized, double-blind trial: 3445 patients with symptomatic Heart Failure and a preserved left ventricular Ejection Fraction (HFpEF) of 45% or more were assigned to receive either spironolactone[1] ($15mg$ to $45mg$ daily) or placebo, with a mean follow-up of 3.3 years. Eligible patients had a history of hospitalization within the previous 12 months, with management of heart failure a major component of the care provided, or an elevated natriuretic peptide level within 60 days before randomization[2]. Trial randomization was stratified according to whether patients were

[1]Spironolactone belongs to the class of mineralocorticoid-receptor antagonists, which have been shown to improve the prognosis for patients with HFpEF [3].

[2]Brain Natriuretic Peptide (BNP) $\geq$ $100\ pg/ml$ or N-terminal pro-BNP (NTproBNP) $\geq$ $360\ pg/ml$.

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

enrolled on the basis of the former criterion (i.e. hospitalization stratum) or the latter criterion (i.e. the BNP stratum). The TOPCAT clinical trial was funded by the National Heart, Lung, and Blood Institute [4], [5]. Details of the trial (inclusion and exclusion criteria and covariates) and its first analysis can be found in [3]. The goal is an evaluation of the effects of spironolactone in patients with HFpEF: the primary outcome is a composite of death from cardiovascular causes, aborted cardiac arrest, or hospitalization for the management of heart failure.

The trial was carried out on a geographically diverse populations with different demographic characteristics (Western vs Eastern regions) [6]. Moreover, the clinical diagnostic criteria were not uniformly interpreted or applied [7]. While country-specific and regional heterogeneity could be viewed as expected statistical variation in a large multinational trial, the differences in patient characteristics, lower event rates, lack of certain drug side effects, and complete lack of treatment effect in Russia and Georgia compared with the Americas lead to a strong suspicion that this data is unreliable in both diagnosis and adherence to treatment [8]. Transportability methods detect significant regional differences and find that measured participant characteristics do not explain the discrepancies between the sites [9]. These methods show promise in detecting anomalies within multi-center Randomized Controlled Trials (RCTs) and in assisting data monitoring to attribute important treatment-effect heterogeneities to participant differences or to site performance differences [10].

Traditional methods based on correlation and survival analysis are prone to confounding and struggle to capture complex interactions: causal discovery is a necessary step in overcoming these issues. Causality can offer new insights into clinical trials: the causal methods described in [11]–[14] show promise for analyzing results from RCTs, via target trials, transportability, and prediction invariance. This paper shows that causal discovery allows us to understand the causal pathways of the main outcomes of the TOPCAT trial, by deriving a causal diagram that provides insight into the regional transportability of the trial results. We were able to demonstrate regional differences of the causal mechanisms between the two main geographical subsets of the trial, offering evidence of the controversial discrepancy of the original study (Western vs Eastern countries). Finally we show that the identified causal models can be linked to a more traditional survival analysis of the trial data, achieving better explanatory performance.

The rest of the paper is organized as follows: the data preparation and the causal discovery are detailed in § 2, and the modelling of clinical outcomes following a survival analysis with the newly found causal relationships is detailed in § 3. A list of acronyms can be found on page 9.

2 Causal Discovery

A total of 3445 participants were recruited at 233 sites in 6 countries (1151 participants in the United States, 326 in Canada, 167 in Brazil, 123 in Argentina, 1066 in Russia, and 612 in Georgia). The original data has a number of time varying variables based on each patient's hospital visits. The original trial data consist of 55 separate datasets with a total of 759 columns. A number of these columns are identifiers such as patient ID, date of observation, etc., that are repeated across datasets. The time and patient coverage of these datasets is varied and we excluded echo-gram data due to its low coverage. There are 7 measured outcomes and the primary outcome is a composite of 3 of them. The outcomes are presented in a time to event format consistent with the survival analysis approach taken in the original paper [3].

For our analysis we converted this data to a time series format to better capture the time varying nature of the data. We aligned the data in time and by patient to create a time series per individual for each variable present in the data. The time resolution was set to 6 months based on a similar average frequency of hospital visits. In each 6 month period starting from initial randomization to 6 years, for each patient we recorded the latest observed data and generated binary variables indicating whether a measured outcome had occurred in the given period or previously. Baseline or static covariates were forward filled from initial randomization. Censored data is automatically handled by this representation, as any period of survival offers a contribution to the data and the time resolution is sufficiently granular to make intra-period dropout approximately equivalent to exact time censoring. By weighting each individual equally as opposed to each row of data we also can avoid underweighting individuals with shorter survival times.

The goal of causal discovery is to identify potential causal drivers from observational data. These methods distinguish correlated variables into causal drivers and confounded variables: the latter are

correlated to the target, but only because there exists a third variable, which influences both. Directed Acyclic Graphs (DAGs) are used to map all a priori assumptions surrounding a causal question and to graphically describe the underlying data generating process. Each node represents a random variable, and directed causal paths are represented by directed edges. The causal graph structure thus provides qualitative information about the conditional independence of the variables of interest, and illustrates potential sources of confounding and selection bias.

Given the complexity of the TOPCAT dataset and the importance of a few clinical outcomes, it would be inefficient to estimate the entire underlying DAG. Instead, we chose a feature selection strategy aimed at deriving the smallest subset of variables, excluding those that do not provide additional information on the outcomes of interest. If the lack of additional information is assessed through conditional independence tests, the problem of feature selection is analogous to identifying the Markov Blanket (MB) of the outcome variable.

Definition 1 *The Markov Blanket (MB) of the outcome Y, $MB(Y)$, is the minimal subset of X, conditioned on which, all other variables of X not included in $MB(Y)$ are independent of Y:*

$$\forall V \in X \backslash MB(Y): \quad \mathbb{P}(Y|MB(Y), V) = \mathbb{P}(Y|MB(Y)) \tag{1}$$

If the prediction model for Y can estimate the underlying true probabilities $\mathbb{P}(Y|MB(Y))$, then the variables included in the MB of the outcome Y are the only variables needed for an optimal prediction in terms of calibration [14].

In order to derive the MB of the primary outcomes, we used PCMCI, specifically designed for time series data, which tests the variables for conditional independence for identified conditioning sets [15]. This method can work directly on the trial data converted to the time series format. PCMCI performs two steps, consisting of an implementation of the PC algorithm, followed by the MCI algorithm. The PC Algorithm identifies a set of potential parents for both each feature and the target. After the potential parents for each variable and the target have been identified, the MCI Algorithm calculates conditional independencies conditioned on those parents. This implementation of PCMCI uses time in order to identify the causal direction.

Our causal discovery is enhanced by the integration of medical context as a constraint structure that our algorithms have to respect. This guarantees both a better performance and time complexity. We specified 8 tiers of variables, i.e. the structure of directionality that we want to enforce in our causal discovery in the form of domain knowledge for groups of variables (cf. § A). The method used for conditional independence tests is the Spearman rank-order correlation coefficient, i.e. a nonparametric measure of the monotonicity of the relationship between two datasets. Unlike the Pearson correlation, the Spearman correlation does not assume that both datasets are normally distributed and that their relationship is linear[3]. We set the threshold of significance to 0.01 for the PC step and to 0.005 for the MCI step of the algorithm (i.e. the threshold for accepting a variable as *causal* in the resulting DAG).

2.1 Causal graphs with domain knowledge

We applied PCMCI to achieve the causal discovery of the MB of the outcomes at three scales:

1. Global level: the whole dataset (cf. Figures 2 and 3).
2. Regional level: the Western countries (US, Canada, Brazil, Argentina) (cf. Figures 4 and 5).
3. Regional level: the Eastern countries (Russia, Georgia) (cf. Figures 6 and 7).

We present those DAGs in the supplementary material (cf. § A). We also show simplified causal graphs for each scale for visualisation purposes, allowing us to focus on the identified relationships of most interest and impact by specifying a restrictive causality threshold to enforce sparsity. These diagrams visualize the causal pathways through which covariates affect outcomes through time. We introduced additional domain knowledge of known causal pathways as edges from treatment to the known side effects and from the constituents of the primary outcome to the primary outcome event itself.

In order to explain the treatment effect, we consider the p-values of the edges that are outgoing from treatment and incoming to the outcomes (cf. Table 1). We notice a correspondence between the

[3]The Spearman rank-order correlation coefficient varies between -1 and $+1$ with 0 implying no correlation. Correlations of -1 or $+1$ imply a perfect monotone increasing and decreasing relationship, respectively.

global and the Western tables, both in terms of significance and of the score signs, as opposed to the Eastern table. [4]. A negative score means that the treatment has the effect of reducing the probability of the outcomes occurring in a time period.

	GLOBAL		WEST		EAST	
Outbound Edge	**Score**	**P-Value**	**Score**	**P-Value**	**Score**	**P-Value**
Primary outcomes						
Outcome cvd death	-0.0379	< 1e-04	-0.0339	2.9997e-04	0.0432	< 1e-04
Outcome hfhosp	0.0353	< 1e-04	-0.0250	7.6480e-03	-0.0041	6.3089e-01
Outcome abortedca	0.1605	< 1e-04	-0.0281	2.7082e-03	0.0794	< 1e-04
Secondary outcomes						
Outcome death	-0.0827	< 1e-04	-0.0502	< 1e-04	0.0706	< 1e-04
Outcome anyhosp	-0.0015	8.1563e-01	-0.0007	9.4204e-01	-0.0032	7.1122e-01
Outcome mi	-0.1182	< 1e-04	-0.0413	< 1e-04	-0.1416	< 1e-04
Outcome stroke	-0.1031	< 1e-04	-0.0750	< 1e-04	-0.0370	< 1e-04

Table 1: Score and p-value of outbound edges of the node *treatment*.

Using Cox regression, the treatment was only significantly associated with Outcome hfhosp in the global dataset (for results of the original study see § A.2). By modelling the causal dynamics and adjusting for the broader MB set, our global analysis has mitigated the regional discrepancies to a greater degree, allowing us to identify significant causal effects of the treatment on the main outcomes as well as on additional relevant outcomes on the global data. (cf. Table 1).

In the global dataset we detected a positive causal effect of treatment on the risk of Outcome cvd death. We noticed a negative causal effect of treatment on the other primary outcomes (Outcome hfhosp and Outcome abortedca). This discrepancy with the Western subset might be a consequence of the conditioning set in the causal discovery that needs to be investigated and addressed. We detected a positive effect of treatment on most secondary outcomes.
In the Western subset we detected a positive causal effect of treatment on almost all the considered primary and secondary outcomes.
In the Eastern subset we noticed fewer significant causal drivers in general, and more discrepancies in the signs of the interactions.

2.2 Bootstrap analysis

We propose a bootstrap analysis as a baseline for evaluating the regional heterogeneities, by comparison with random partitions of the data. When examining the observed differences in causal scores and graph structures across regions, it is important to contextualize these numbers. With random variation in the data some level of difference would be expected and we can quantify this expected variation using bootstrap sampling. By drawing random samples of data partitions that are uncorrelated to Region or other covariates, we can observe the expected distribution of the differences due to randomness and compare them to the observed regional behavior. We drew 100 of these bootstrap samples, randomly partitioning the data into two parts. After running causal discovery on all the sets of subsections, we calculated a number of metrics to capture the variation in the identified causal structures in Table 2.

Histograms of SHD and the number of contradictions can be found in Figure 1, and details of the differences among single outcomes can be found in Table 3.

Looking at a broader causal picture has allowed us to capture regional differences very strongly when a univariate survival analysis could not. We can see in Tables 2 and 3 that stratifying the data by region led to significantly higher differences in results compared to random partitions. Differences in score (SMAPE) were significant with regard to treatment (86% quantile) and extreme when taking into account other covariates (100% quantile). Measurements around graph structure (SHD) also gave a clear picture (97% quantile). The number of contradictory causal conclusions between regions (26) was higher than any observed bootstrap sample with the highest observed being 19 (100%

[4]The p-value and Causal Score are the result of the conditional independence test carried out by PCMCI with a significance threshold of 0.01.

[5]Symmetric Mean Absolute Percentage Error (SMAPE).

Measure	SMAPE Treatment	SMAPE All	SHD	Contradictions
Observed Regional Measure	1.500	1.560	169	26
Observed quantile	86%	100%	97%	100%
Bootstrap 25% quantile	1.044	1.532	112	7
Bootstrap 50% quantile	1.223	1.536	129.5	9.5
Bootstrap 75% quantile	1.401	1.541	144	12

Table 2: Comparison of Observed Regional Score with Bootstrap Distribution.

SMAPE[5]Treatment: SMAPE between causal scores of the treatment on each outcome.
SMAPE All: SMAPE between the causal scores of every covariate on each outcome.
Structural Hamming Distance (SHD): between the graphs discovered on each subset.
Contradictions: the number of existing causal edges with opposite signs.

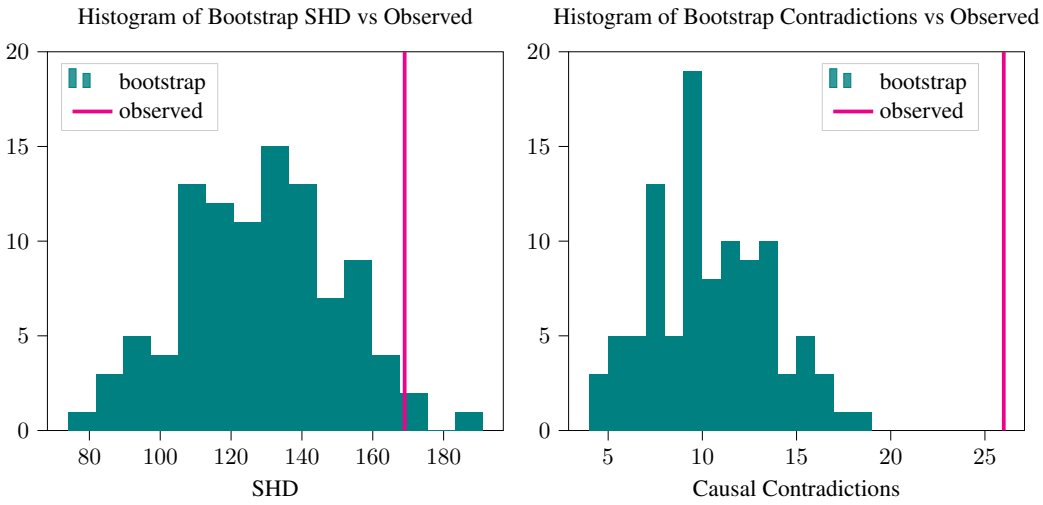

Figure 1: Bootstrap histograms: SHD on the left, contradictions on the right.

Outcome	Treatment Difference	Quantile	Mean Difference	Quantile
Outcome cvd death	0.077	57%	0.086	99%
Outcome abortedca	0.107	53%	0.077	96%
Outcome hfhosp	0.021	52%	0.071	100%
Outcome death	0.121	80%	0.083	98%
Outcome anyhosp	0.002	11%	0.038	100%
Outcome mi	0.100	79%	0.070	100%
Outcome stroke	0.038	30%	0.056	99%

Table 3: Comparison of Observed Regional Score with Bootstrap Distribution per Outcome.

quantile). This comparison strongly supports the claim of regional treatment discrepancy in terms of the difference between the corresponding causal structures.

3 Survival Analysis

We propose the use of causal discovery as an automatic tool for feature selection of the most influential covariates for a traditional survival analysis. Generally, covariates are selected manually or using correlation methods, without taking into account the causal relationship to the outcomes of interest. In RCTs with a survival outcome, treatment effects are generally estimated with the Cox hazard ratio, often without adjusting for known prognostic variables. When interpreting Cox regression through the lens of causality, omission of confounding terms creates problems, despite randomization of treatment at the origin of time ($t = 0$). Therefore, the hazard ratio does not admit a causal interpretation in the case of unknown heterogeneity [16], as the risk sets beyond the first event time are comprised of the subset of individuals who have not previously failed. Despite randomization at $t = 0$, the contributions to the Cox partial likelihood following the first failure will not remain randomized.

On the other hand, the effect of treatment in terms of an Accelerated Failure Time (AFT) model, does yield a causal measure of treatment effects as the hazard function has an additive form in the covariates [16] (cf. § A.3 for more details). The coefficient of the treatment variable in the AFT model can be interpreted as slowing down or speeding up the survival function. A coefficient b changes the expected survival time by a factor of $\exp(b)$. We can see in the tables below that the Global and Western AFT models identify a significant effect of the treatment on heart failure hospitalization. This coefficient of 0.28 is interpreted as meaning that the expected survival time before heart failure hospitalization is $\exp(0.28) - 1 = 32.3\%$ longer given the treatment than without.

We present results of goodness of fit for the AFT regression of the outcomes in the form of Bayesian Information Criterion (BIC). For large datasets, the model with minimum Bayesian Information Criterion (BIC) ideally corresponds to the candidate model which is a posteriori most probable [17]. [6] Results of the AFT regression model for the global dataset and the Western and Eastern subsets can be found in Table 4. The rows correspond to the particular predicted outcome, whereas the columns correspond to two feature selection strategies before AFT modeling: (1) Features corresponding to the MB of the outcomes (cf. Section 2); (2) Features used in the primary study in [3]. We can see that the MB yielded by causal discovery in all cases leads to a lower BIC and correspondingly a better model of the data. This accounts for increased complexity in the model with the introduction of the new MB covariates. The absolute scale of the BIC for each model is a function of the size of the data so does not allow for relative comparison across regions but by looking at the baseline *Original model* on each region we can see relatively how well we capture the survival process in this data.

Outcome	Coef MB	p-val MB	BIC MB	Coef Orig	p-val Orig	BIC Orig	
hfhosp	0.28	1.87e-02	3349.36	0.29	2.26e-02	3512.87	**GLOBAL**
abortedca	0.16	5.16e-01	154.17	0.14	5.67e-01	163.89	
cvd death	0.10	3.67e-01	2997.92	0.11	3.33e-01	3040.80	
Total			6501.45			6717.56	
hfhosp	0.28	4.34e-02	2601.16	0.30	3.86e-02	2673.64	**WEST**
abortedca	0.24	4.25e-01	106.81	0.24	4.47e-01	113.83	
cvd death	0.40	5.32e-03	1840.00	0.39	7.20e-03	1872.07	
Total			4547.97			4659.54	
hfhosp	0.36	1.57e-01	535.49	0.26	3.53e-01	609.27	**EAST**
abortedca	-0.02	9.56e-01	53.80	-0.02	9.62e-01	58.99	
cvd death	-0.22	2.30e-01	1104.85	-0.22	2.21e-01	1149.90	
Total			1694.14			1818.16	

Table 4: Coefficients with corresponding p-values and BICs of AFT models.

4 Conclusion and Perspectives

Complex trials with heterogeneities in the data pose a significant challenge for traditional techniques of statistical analysis. Suboptimal patient cohort selection and recruitment, paired with ineffective monitoring of patients, are among the main causes for high trial failure rates in phases 2 and 3 [1], [2]. Our ambition is to use causality in order to provide earlier determination of drug futility, inform adaptive trial design, and shorten overall drug development time. In this paper we show that cutting edge techniques of causal discovery augmented with human domain knowledge can offer novel insights and achieve results at greater level of significance.

Traditional correlational methods are limited by confounding and multiple pathways and traditional survival modelling is restricted to a univariate outcome, necessitating the definition of a single primary outcome. Causal discovery allows us to understand the interaction between outcomes and covariates leading to a diagram describing the causal process of the TOPCAT dataset. Our causal analysis has mitigated the regional discrepancies, allowing us to identify significant global causal effects of the treatment on the main outcomes as well as on additional relevant outcomes. After derivation of the regional causal diagrams, our bootstrap analysis provided evidence for a different causal effect of the

[6] Bayesian Information Criterion (BIC) is based on the empirical log-likelihood L and does not require the specification of priors. The BIC for candidate model M_k is defined as $\text{BIC} = -2 \ln L(\theta_k|y) + k \ln n$ where k is the number of parameters estimated by the model and n is the sample size of observations y.

treatment by region. We also applied an AFT modelling to the data and were able to achieve lower BIC after MB feature selection.

As a future extension of our work, we intend to achieve better causal effect estimation and design of clinical trials, as secondary relationships are learned from this dataset via full causal discovery. A combination of more in depth domain knowledge and data driven causal inference methods could lead to better survival models. Additional covariates could be added thus accounting for further potential confounding. The causal framework also allows conclusions from trials to go beyond binary outcomes and further refine our analysis (e.g. including multiple levels/doses of treatment). These conclusions could be continuously derived during a clinical trial, potentially leading to changes in its framework or to an early termination.

Transportability of causal graphs [18]–[22] refers to the problem of extrapolating experimental findings across domains (i.e., settings, populations, environments) that differ both in their distributions and in their inherent causal characteristics [23], [24]. For instance, this can allow us to infer the causal effect at one population from experiments conducted in a different population after noticing that some of the covariate distributions are different. The subject of transportability in epidemiology is also related to data fusion, i.e. the combination of compatible graphs according to their degree of confidence, similarly to [25]. Since the TOPCAT study combined data from multiple geographical sources that had different underlying dynamics leading to mixed results, we could combine compatible graphs from different sub-populations according to their degree of confidence. We could similarly validate data sources and our confidence in their adherence to the study design by comparing their causal dynamics to known ground truths.

References

[1] S. Harrer, P. Shah, B. Antony, and J. Hu, "Artificial intelligence for clinical trial design," *Trends in pharmacological sciences*, vol. 40, no. 8, pp. 577–591, 2019.

[2] G. A. Van Norman, "Phase II trials in drug development and adaptive trial design," *JACC: Basic to Translational Science*, vol. 4, no. 3, pp. 428–437, 2019.

[3] B. Pitt, M. A. Pfeffer, S. F. Assmann, *et al.*, "Spironolactone for heart failure with preserved ejection fraction," *New England Journal of Medicine*, vol. 370, pp. 1383–1392, 2014.

[4] L. National Heart, B. Institute, *et al.*, "Aldosterone antagonist therapy for adults with heart failure and preserved systolic function (TOPCAT)," *ClinicalTrials. gov Web site. http://www. clinicaltrials. gov/ct2/show/NCT00094302*, 2009.

[5] *Treatment of preserved cardiac function heart failure with an aldosterone antagonist - TOPCAT*, `https://www.acc.org/latest-in-cardiology/clinical-trials/2014/06/04/12/40/topcat`.

[6] M. A. Pfeffer, B. Claggett, S. F. Assmann, *et al.*, "Regional variation in patients and outcomes in the treatment of preserved cardiac function heart failure with an aldosterone antagonist (TOPCAT) trial," *Circulation*, vol. 131, no. 1, pp. 34–42, 2015.

[7] W. Aronov, "Lessons learned from the TOPCAT trial," *J. Clin. Exp. Cardiol*, vol. 6, 2014.

[8] M. R. Bristow, J. Silva Enciso, B. J. Gersh, *et al.*, "Detection and management of geographic disparities in the TOPCAT trial: Lessons learned and derivative recommendations," *JACC: Basic to Translational Science*, vol. 1, no. 3, pp. 180–189, 2016.

[9] S. A. Berkowitz, K. E. Rudolph, and S. Basu, "Detecting anomalies among practice sites within multicenter trials: An application of transportability methods to the TOPCAT and accord BP trials," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 3, e004907, 2019.

[10] K. E. Rudolph, J. Levy, N. M. Schmidt, E. Stuart, and J. Ahern, "Using transportability to understand differences in mediation mechanisms across trial sites of a housing voucher experiment," *Epidemiology (Cambridge, Mass.)*, vol. 31, no. 4, p. 523, 2020.

[11] J. Q. Wu, N. Horeweg, M. de Bruyn, *et al.*, "Automated causal inference in application to randomized controlled clinical trials," *Nature Machine Intelligence*, vol. 4, no. 5, pp. 436–444, 2022.

[12] I. J. Dahabreh, S. E. Robertson, J. A. Steingrimsson, E. A. Stuart, and M. A. Hernan, "Extending inferences from a randomized trial to a new target population," *Statistics in medicine*, vol. 39, no. 14, pp. 1999–2014, 2020.

[13] M. Prosperi, Y. Guo, M. Sperrin, *et al.*, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nature Machine Intelligence*, vol. 2, no. 7, pp. 369–375, 2020.

[14] M. Piccininni, S. Konigorski, J. L. Rohmann, and T. Kurth, "Directed acyclic graphs and causal thinking in clinical risk prediction modeling," *BMC medical research methodology*, vol. 20, no. 1, pp. 1–9, 2020.

[15] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science advances*, vol. 5, no. 11, eaau4996, 2019.

[16] O. O. Aalen, R. J. Cook, and K. Røysland, "Does Cox analysis of a randomized survival study yield a causal treatment effect?" *Lifetime data analysis*, vol. 21, no. 4, pp. 579–593, 2015.

[17] A. A. Neath and J. E. Cavanaugh, "The bayesian information criterion: Background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012.

[18] E. Bareinboim and J. Pearl, "Transportability of causal effects: Completeness results," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, 2012, pp. 698–704.

[19] E. Bareinboim and J. Pearl, "Causal transportability with limited experiments," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[20] E. Bareinboim and J. Pearl, "Meta-transportability of causal effects: A formal approach," in *Artificial Intelligence and Statistics*, PMLR, 2013, pp. 135–143.

[21] M. L. Petersen, "Compound treatments, transportability, and the structural causal model: The power and simplicity of causal graphs," *Epidemiology*, vol. 22, no. 3, pp. 378–381, 2011.

[22] S. Lee, J. D. Correa, and E. Bareinboim, "Generalized transportability: Synthesis of experiments from heterogeneous domains," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, AAAI Press New York, NY, USA, 2020.

[23] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[24] C. R. Lesko, A. L. Buchanan, D. Westreich, J. K. Edwards, M. G. Hudgens, and S. R. Cole, "Generalizing study results: A potential outcomes perspective," *Epidemiology (Cambridge, Mass.)*, vol. 28, no. 4, p. 553, 2017.

[25] D. Alrajeh, H. Chockler, and J. Y. Halpern, "Combining experts' causal judgments," *Artificial Intelligence*, vol. 288, p. 103 355, 2020.

A Supplementary Material

Acronyms

AFT Accelerated Failure Time. 6, 7, 12

BIC Bayesian Information Criterion. 6, 7

DAG Directed Acyclic Graph. 3, 13

HFpEF Heart Failure and a preserved left ventricular Ejection Fraction. 1, 2

MB Markov Blanket. 3, 4, 6, 7

Outcome abortedca Aborted cardiac arrest - primary outcome. 4, 5, 11

Outcome anyhosp Hospitalisation for any reason. 4, 5, 11

Outcome cvd death Death from cardiovascular causes - primary outcome. 4, 10, 11

Outcome death Death from any cause. 4, 5, 11

Outcome hfhosp Hospitalization for hearth failure - primary outcome. 4, 5, 11

Outcome mi Myocardial infarction. 4, 5, 11

Outcome stroke Stroke. 4, 5, 11

RCT Randomized Controlled Trial. 2, 5

SHD Structural Hamming Distance. 4, 5

SMAPE Symmetric Mean Absolute Percentage Error. 4, 5

TOPCAT Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist. 1–3, 6, 7

A.1 Tiers for causal discovery

```
[

[ ‘Americas’, ‘Region interaction’, ‘treatment’, ‘age entry’, ],

[ ‘blood potassium’, ‘systolic blood pressure’, ‘diastolic blood
pressure’, ‘Left ventricular ejection fraction’, ‘Heart Failure History’,
‘Previous Atrial Fibrillation’, ‘Known chronic hepatic disease’, ‘Heart
Failure hospitalization’, ‘Angina’, ‘Chronic Obstructive Pulmonary
Disease’, ‘Asthma’, ‘Peripheral Arterial Disease’, ‘Dyslipidemia’,
‘Implanted cardioverter defibrillator’, ‘Pacemaker implanted’, ‘Atrial
fibrillation’, ‘Coronary artery bypass graft surgery’, ‘Percutaneous
Coronary Revascularization’, ‘Stroke’, ‘myocardial infarction’, ‘smoker’,
‘years smoker’, ‘previous smoker’, ‘cigarettes smoke a day’, ‘sodium
intake score’, ‘alcohol per week’, ‘activity level’, ‘weight’, ‘waist
circumference’, ‘heart rate’, ‘Use of blood pressure lowering medication’,
‘Use of Non-cardiovascular medication’, ‘Use of cardiac medications’,
‘aspirin use’, ‘nitrate use’, ‘statin use’, ‘warfarin use’, ‘diuretic use’,
‘Use of Angiotensin receptor blocker’, ‘Use of Angiotensin converting enzyme
inhibitor’, ‘white blood cell count’, ‘hemoglobin’, ‘Platelet’, ‘Albumin’,
‘Total Bilirubin’, ‘blood glucose’, ‘Alkaline Phosphatase’, ‘Aspartate
Aminotransferase’, ‘blood sodium’, ‘blood chloride’, ‘creatinine result’,
‘gfr’, ‘Insulin treated’, ‘Not Insulin treated’, ‘IN HOSP’, ‘left atrial
volume 4’, ‘left atrial volume 2’, ],

[ ‘serum potassium high’, ‘Systolic blood pressure (SBP) > 160 mm Hg’, ],

[ ‘NYHA class’, ‘hyperkalemia’, ‘Hypertension’, ],

[ ‘creatinine doubled’, ‘hyperkalemia hospitalization’, ],

[ ‘Outcome anyhosp’, ‘Outcome mi’, ‘Outcome stroke’, ‘Outcome abortedca’,
‘Outcome hfhosp’, ],

[ ‘Outcome death’, ‘Outcome cvd death’, ‘hyperkalemia side effect
discontinued’, ‘potassium side effect discontinued’, ‘renal function side
effect discontinued’, ],

[‘Outcome primary ep’, ‘target’],

]
```

A.2 Comparison with the original study

For the sake of comparison with the primary study in [3], we reproduce their original table of results:

Outcome	**Hazard Ratio with Spironolactone** (95% **CI)**	**P-Value**
Primary outcome	0.89 (0.77–1.04)	0.14
Primary outcomes		
Outcome Outcome cvd death	0.90 (0.73–1.12)	0.35
Outcome abortedca	0.60 (0.14–2.50)	0.48
Outcome hfhosp	0.83 (0.69–0.99)	0.04
Secondary outcomes		
Outcome death	0.91 (0.77–1.08)	0.29
Outcome anyhosp	0.94 (0.85–1.04)	0.25
Outcome mi	1.00 (0.71–1.42)	0.98
Outcome stroke	0.94 (0.65–1.35)	0.73

Table 5: Unadjusted hazard ratios calculated via Cox proportional-hazards models.

The original study in [3] concluded that adding spironolactone to existing therapy in patients with heart failure and a preserved ejection fraction did not significantly reduce the incidence of the primary outcome (cf. Table 5). As for the single Primary outcomes, spironolactone did not significantly reduce the composite primary end point of death from cardiovascular causes, aborted cardiac arrest, or hospitalization for heart failure. Also, neither the time to first hospitalization for any reason nor the time to death from any cause was significantly altered by random assignment to spironolactone. These two inclusive outcome measures are important for assessing overall risk versus benefit of the drug.

Post hoc analysis indicated marked regional differences in event rates in the Supplementary Appendix of [3] (cf Table 6). In the Americas (the United States, Canada, Brazil, and Argentina), the primary outcome occurred in 242 patients in the spironolactone group (27.3%) and 280 patients in the placebo group (31.8%). In Russia and Georgia, the primary outcome occurred in 78 patients in the spironolactone group (9.3%) and 71 patients in the placebo group (8.4%). However, the pre-specified test for interaction between region and study group was not significant ($p = 0.12$).

	Western subset (USA CAN ARG BRA) (N=1,767)	Eastern subset (RUS GEO) (N=1,678)
Hazard Ratio	0.82 (0.69-0.98)	1.10 (0.79-1.51)
P-Value	0.026	0.576

Table 6: Unadjusted hazard ratios calculated via Cox proportional-hazards models by geographic region.

A.3 Accelerated Failure Time (AFT)

AFT Models assume the following form:

$$\log(T) = -\log(\theta) + \log(T\theta) = -\log(\theta) + \epsilon \tag{2}$$

where T denotes the time to event, ϵ is distributed according to the underlying parametric assumption, and θ denotes the joint effect of covariates x_i (typically $\theta = \boldsymbol{\beta}^T \boldsymbol{x} = \exp\left(\sum_i \beta_i x_i\right)$). This reduces AFT modelling to a regression where we estimate the fixed effect of our covariates with the chosen parametric distribution modelling the noise.

We used a Weibull distribution for our analysis. We can write Equation 2 as

$$\log(T) = Y = \mu + \boldsymbol{\alpha}^T \boldsymbol{x} + \sigma W \tag{3}$$

where $\boldsymbol{x}$ is the set of covariates and W has extreme value distribution. Given transformations

$$\begin{cases} \gamma = 1/\sigma \\ \lambda = \exp\left(-\mu/\sigma\right) \\ \boldsymbol{\beta} = -\boldsymbol{\alpha}/\sigma \end{cases} \tag{4}$$

we obtain a Weibull model with proportional basal hazard

$$\begin{aligned} h(t|\boldsymbol{x}) &= (\gamma \lambda t^{\gamma-1}) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right) = \\ &= h_0(t) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right) \end{aligned} \tag{5}$$

Notice that the logarithm of the hazard in Equation 5 is additive in the different covariates:

$$\log[h(t|\boldsymbol{x})] = \log[h_0(t)] + \boldsymbol{\beta}^T \boldsymbol{x} = \log[h_0(t)] + \sum_i \beta_i x_i \tag{6}$$

A.4 Directed Acyclic Graphs (DAGs)

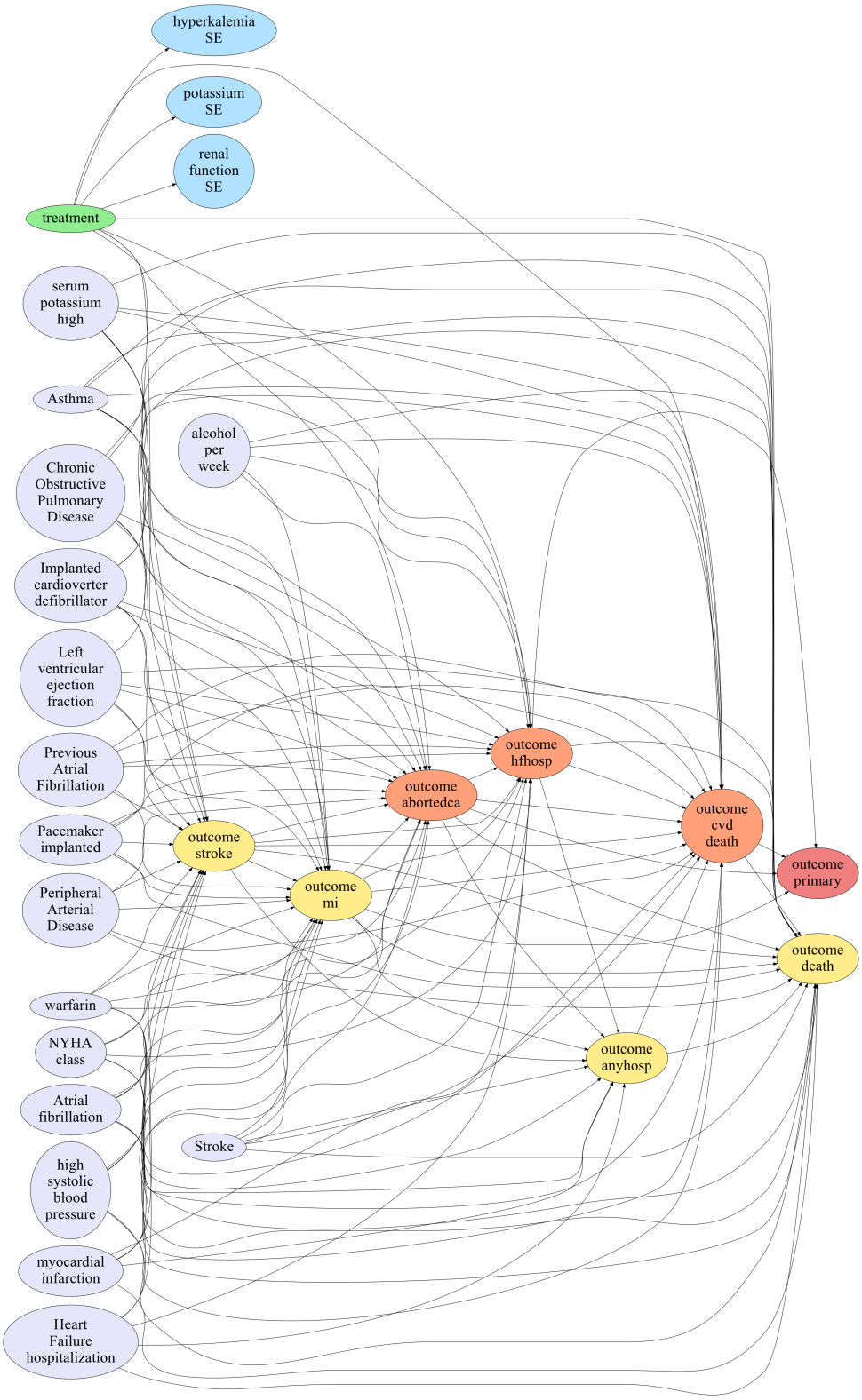

Figure 2: DAG for the global dataset

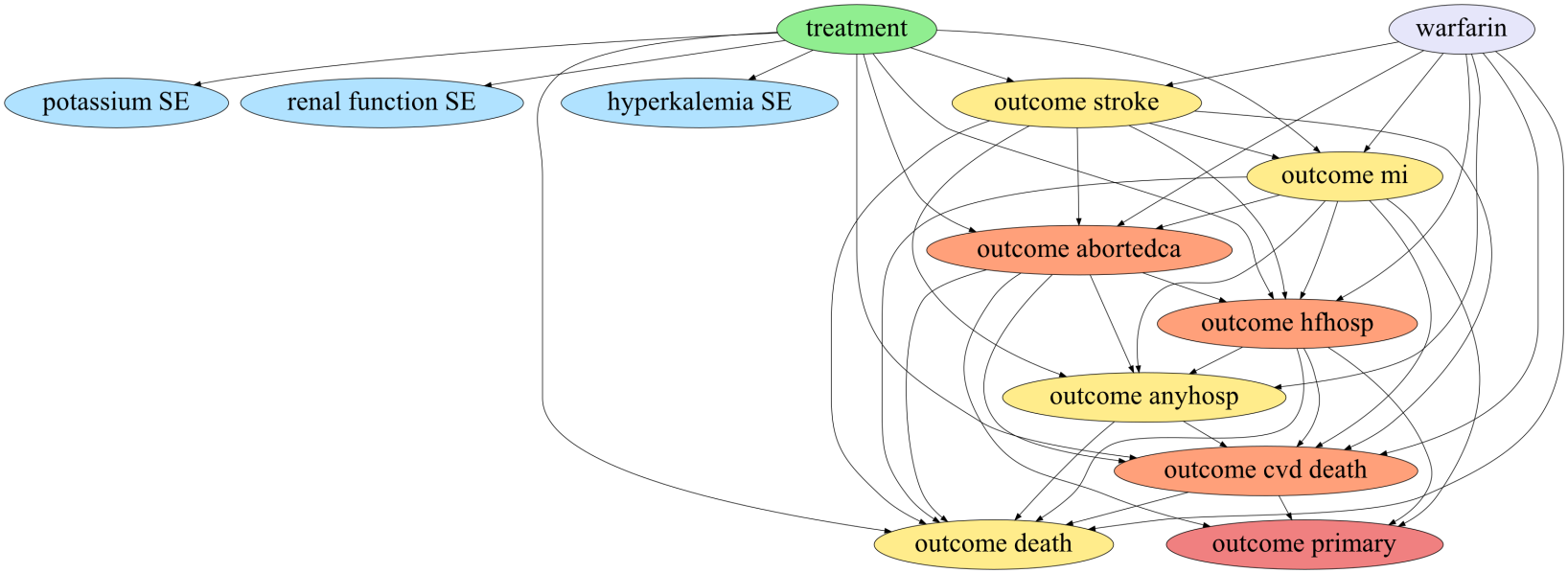

Figure 3: DAG for the global dataset - after pruning:
non outcome nodes with less than 7 outbound edges have been pruned, for better visualization.

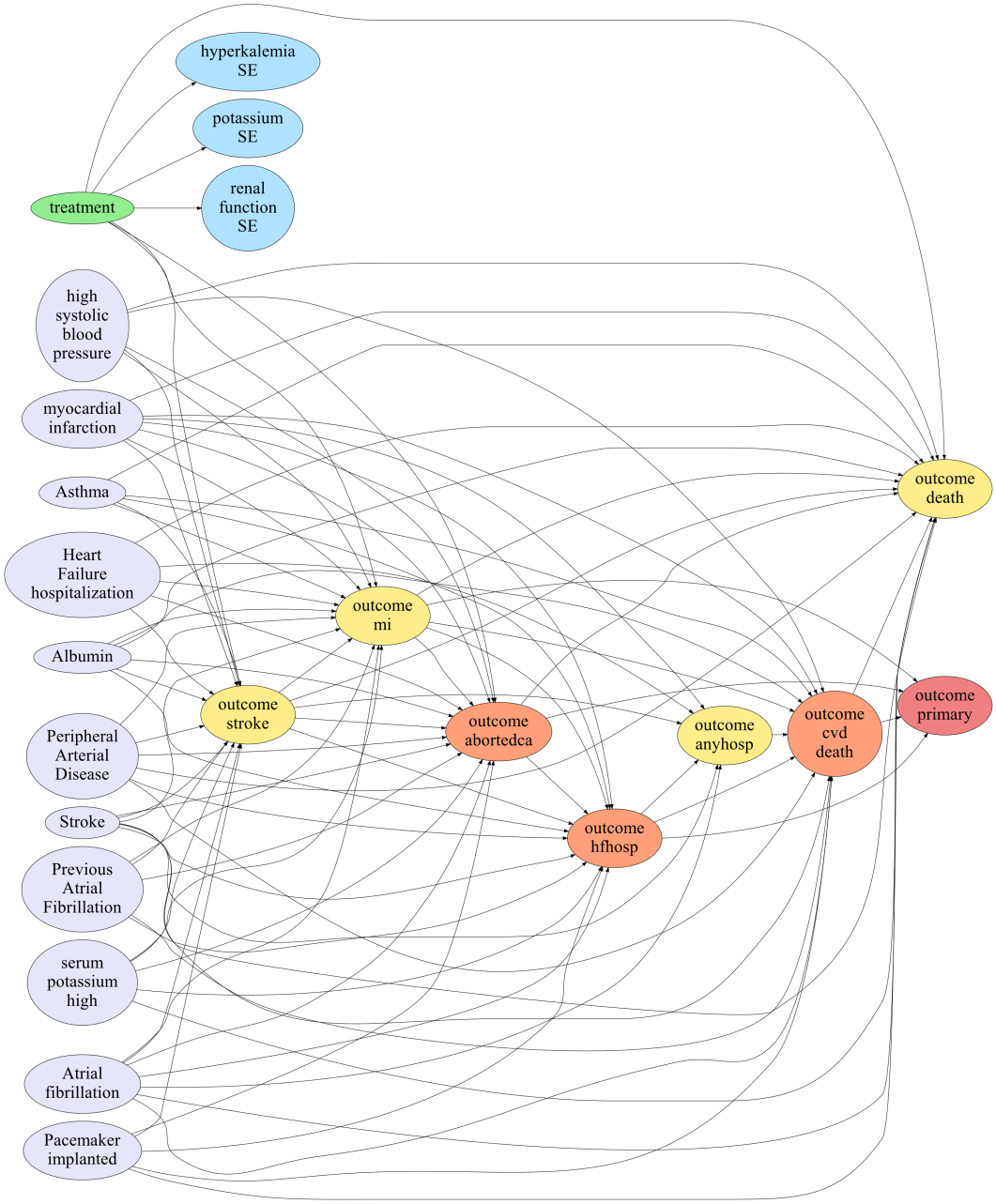

Figure 4: DAG for the Western regional subset

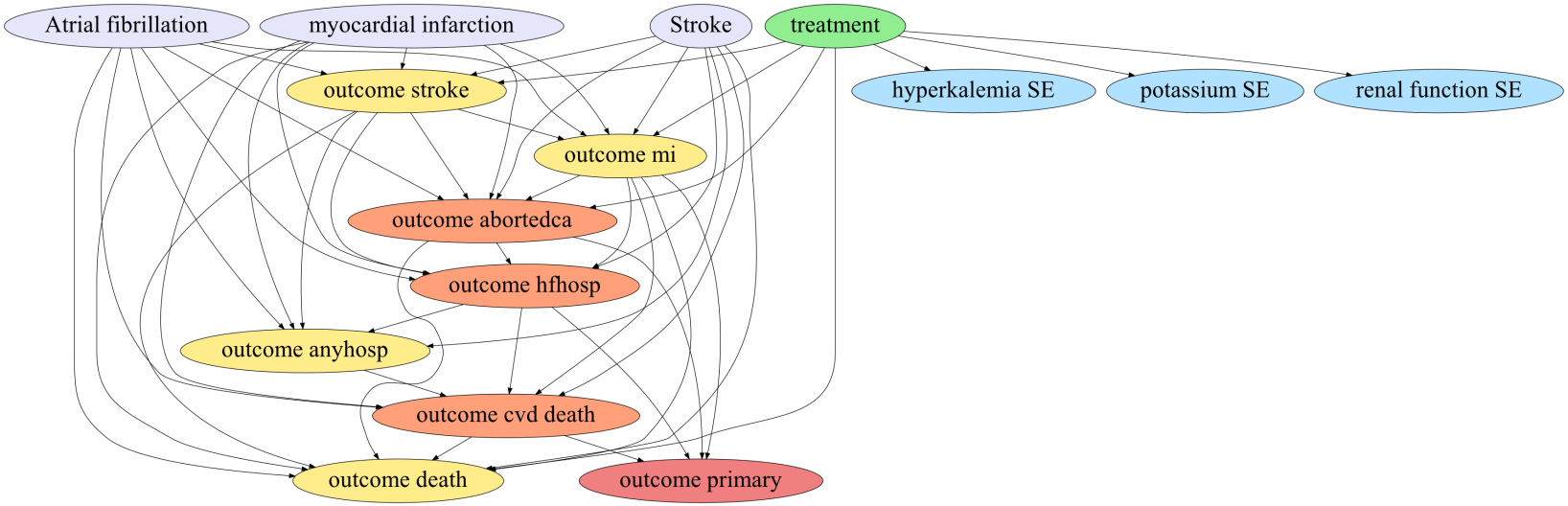

Figure 5: DAG for the Western regional subset - after pruning:
non outcome nodes with less than 7 outbound edges have been pruned.

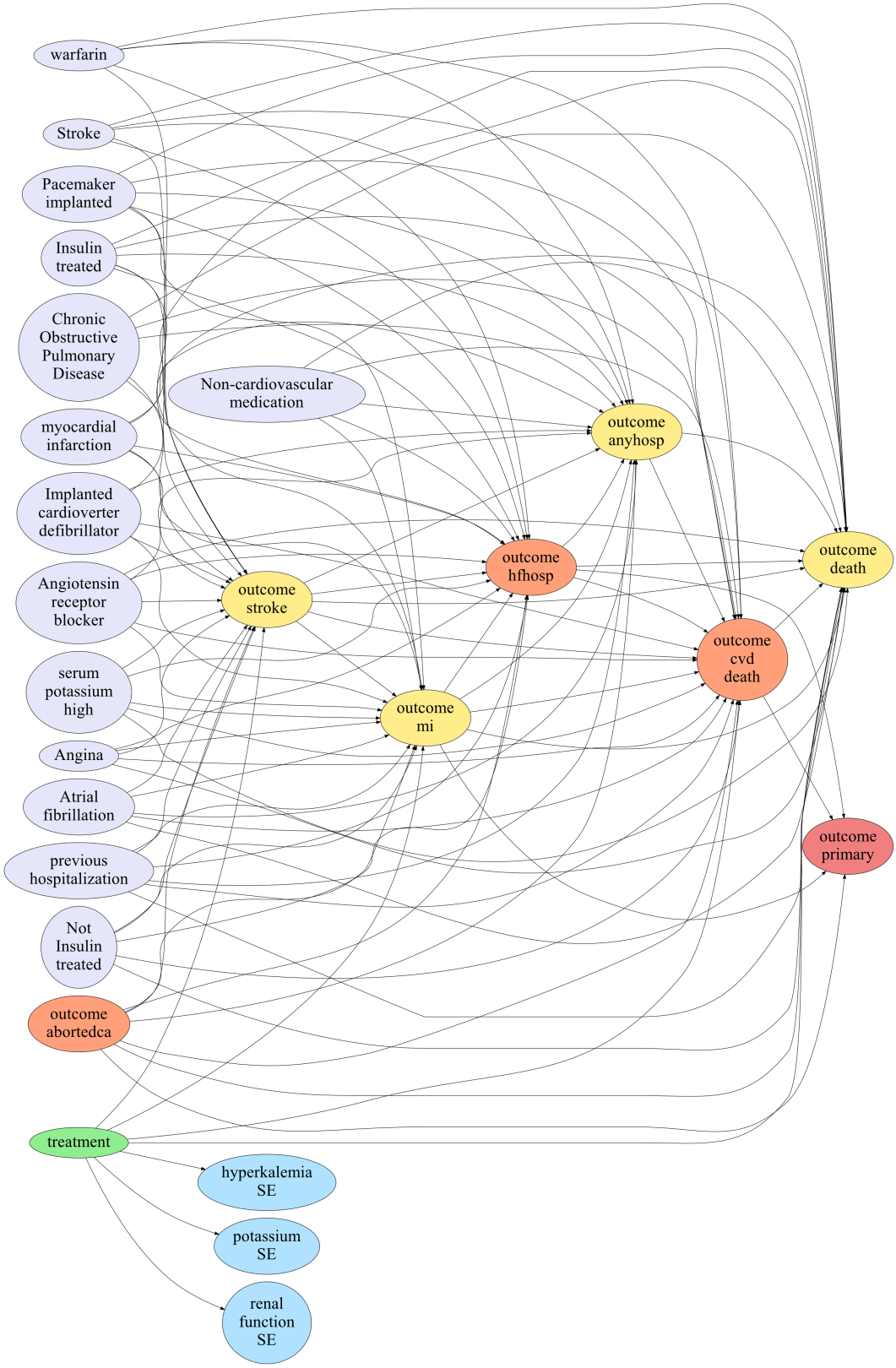

Figure 6: DAG for the Eastern regional subset

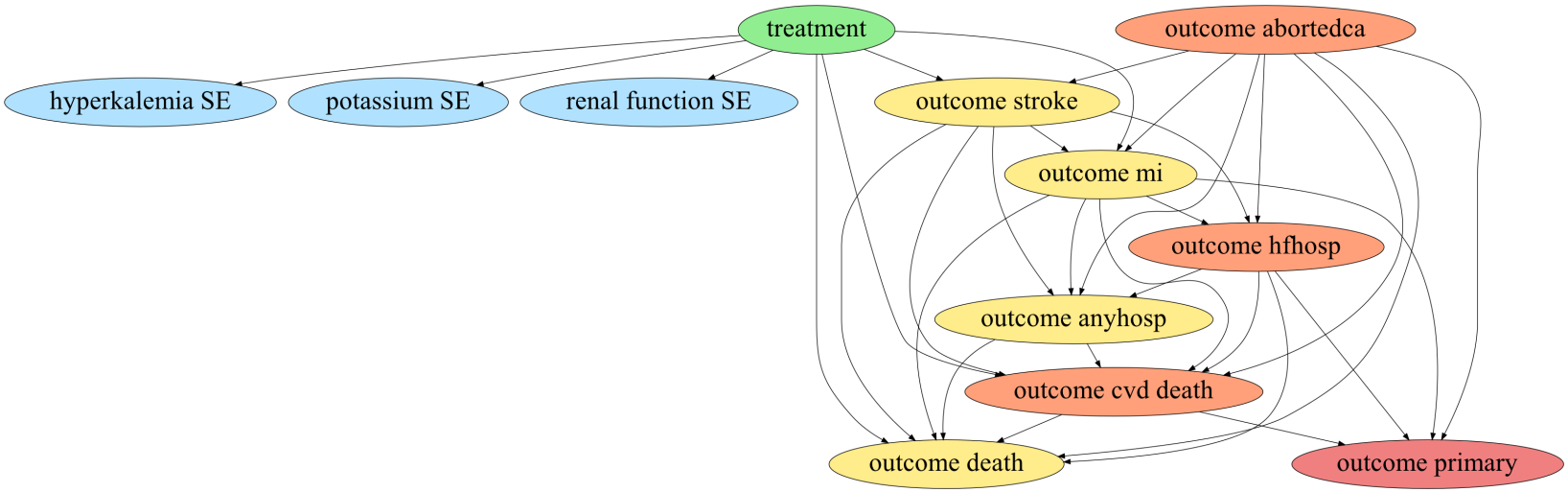

Figure 7: DAG for the Eastern regional subset - after pruning:
non outcome nodes with less than 7 outbound edges have been pruned.