A causal machine learning framework for pharmacovigilance signal detection in electronic health records: Drug-induced acute kidney injury

Stella Dimitsaki STELLA.DIMITSAKI@SORBONNE-UNIVERSITE.FR
LIMICS, Inserm, Université Sorbonne Paris-Nord, Sorbonne Université, Paris, France

Corinne Isnard Bagnis CORINNE.BAGNIS@APHP.FR
Département de Néphrologie, APHP Sorbonne University, Paris, France

Pantelis Natsiavas PNATSIAVAS@CERTH.GR
Centre for Research and Development Hellas, Institute of Applied Biosciences, Thessaloniki, Greece

Marie-Christine Jaulent MARIE-CHRISTINE.JAULENT@INSERM.FR
LIMICS, Inserm, Université Sorbonne Paris-Nord, Sorbonne Université, Paris, France

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

An increasing number of studies in the field of pharmacovigilance have been testing different Artificial intelligence (AI) approaches on Real-World Data (RWD). These studies use conventional AI to predict adverse drug reactions by detecting the correlation of patient characteristics with adverse effects (AE). Typically, these studies do not aim to establish any causal relationship between the suspect drugs and the AE. Causal inference enables machine learning methods to estimate the treatment effect of medical interventions using RWD. However, applying causal inference to RWD presents its own set of challenges. Therefore, a comprehensive framework is essential. In this study, the integration of PRINCIPLED, a process guide for causal inference using healthcare data, electronic health records from the MIMIC-IV database, and Causal Machine Learning (CML) to detect drug-induced acute kidney injury, demonstrates a framework that can provide interpretable, reproducible, and clinically relevant information. The results position CML as a promising approach for improving the accuracy, transparency, and regulatory acceptance of pharmacovigilance systems.

Keywords: Causal machine learning, pharmacovigilance, real-world data

1. Introduction

Pharmacovigilance (PV) studies the causal relationship between a drug and an adverse effect (AE). Traditional PV approaches focus on statistical methods that detect correlations between drugs and AE in PV databases and more recently in real-world data (RWD) (Crisafulli et al., 2023). The term RWD refers to data collected outside the controlled environment of clinical trials, such as electronic health records (EHRs). Machine learning (ML) can identify associations in RWD that traditional statistical methods, due to data complexity, cannot detect, making it a more expressive analytical approach. However, observational studies can be subject to bias as evidence and reproducibility concerns often arise from incomplete documentation of cohort definitions, confounder handling, or code (Wang et al., 2022). Conventional ML makes predictions using historical data, but decision-making in PV requires comparing potential outcomes with and without the suspect drug. Thus, causal machine learning (CML) could offer mechanistic insights to observational studies in PV by

addressing confounding and prove an initial causal association between the drug and the AE. So far, few studies have explored various CML approaches in EHRs for PV. Wang et al. (2023) used targeted maximum likelihood estimation to estimate drug-induced liver injury, and Zhang et al. (2023) applied causal discovery algorithms to EHRs to study remdesivir-related kidney injury.

This study[1] examines whether CML can support PV by applying an existing causal analysis workflow to a drug-induced acute kidney injury (DAKI) case study using EHR data[2]. As responsible use of CML requires protocols, standardized reporting and alignment with regulatory expectations (Feuerriegel et al., 2024), we follow PRINCIPLED (Desai et al., 2024), a process guide for causal inference using healthcare data, and adapt it to our study. Our goal is to evaluate whether this framework can produce clinically plausible and robust results under real-world constraints.

2. Methods

This section maps the steps in the PRINCIPLED (Desai et al., 2024). Figure 1 presents an overview of the framework in the context of our study.

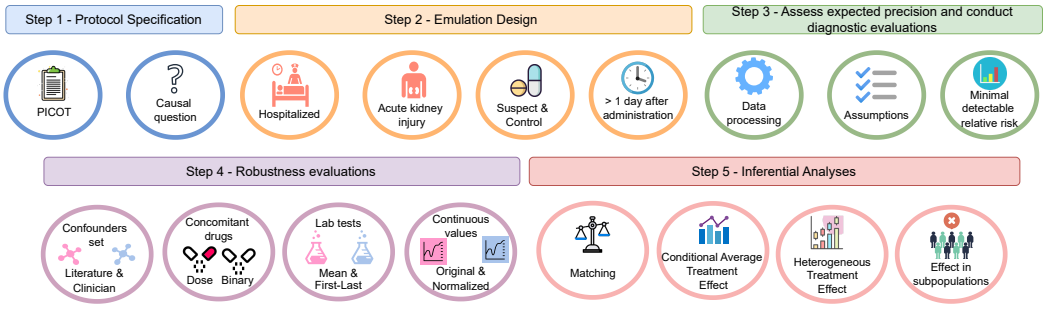

Figure 1: Methodology pipeline according to PRINCIPLED steps

2.1. Step 1: Protocol Specification

2.1.1. CASE STUDY: DRUG-INDUCED ACUTE KIDNEY INJURY (DAKI)

Acute kidney injury (AKI) is defined as a decline in kidney function with increased serum creatinine or reduced urine output within 7 days (Kellum et al., 2012). Mehta et al. (2015) supports that DAKI incident typically occurs within 7 days of drug initiation. AKI affects up to 20% of hospitalized patients (Waikar et al., 2008; Zeng et al., 2014; Susantitaphong et al., 2013), and DAKI accounts for 19–26% of cases (Karimzadeh et al., 2023).

Ryan et al. (2013) provides a reference set of drugs classified as positive (potentially causing DAKI) or negative controls (verified not to cause DAKI). Their study is based on a systematic literature review. They extracted product labels using natural language processing tools and calculated a minimum detectable relative risk (MDRR) to ensure sufficient sample size and power (Armstrong, 1987), resulting in 13 positive (suspect to AKI) and 24 negative controls (not related to AKI).

1. The main section concentrates primarily on accessibility, while the technical details are elaborated in the appendices.
2. Code is available here: https://github.com/Dimstella/PRINCIPLED-CLM-DAKI

More recently, Fernández-Llaneza et al. (2024) compiled 200 positive drugs from multiple sources, including spontaneous reporting systems, drug databases, NephroTox [3], and peer-reviewed literature. Drugs were categorized to very strong, strong, moderate, or limited effect on AKI based on disproportionality analysis, event frequency, and published evidence.

Using these studies, we compiled a list of positive and negative drugs, but not all appear in every EHR system or have enough patients for reliable analysis. Therefore, in our study, we consider the subset of drugs, including all drug synonyms, for which we retrieved enough data in the MIMIC-IV database (Johnson et al., 2020).

2.1.2. FROM TARGET TRIAL PROTOCOL TO CAUSAL QUESTION

The study is framed as an emulation of a "target trial", specifying eligibility, assignment, follow-up, and causal contrast. We applied the PICO (Population, Intervention, Control, Outcome) framework, Guyatt and Rennie (1993), an extension of it that adds Time as a critical component, particularly important for time-series data such as EHRs, as highlighted in Doutreligne et al. (2025). Table 1 describes the PICOT framework for PV signal detection of DAKI.

Table 1: PICOT framework for drug-induced acute kidney injury (DAKI)

PICOT	**Application in DAKI prediction**
Population	Hospitalized patients with AKI stage ≥ 1 based to KDIGO guidelines
Intervention	Ibuprofen and Ketorolac (NSAID, *very strong* effect), Vancomycin (Antibacteria, *strong* effect), Lisinopril (RAAS-acting agent, *strong* effect), Furosemide (Diuretic, *moderate* effect), Pantoprazole and Omeprazole (PPI, *moderate* effect), and Allopurinol (Xanthine oxidase inhibitor, *limited* effect)
Control	Simethicone, Prochlorperazine, and Lactulose
Outcome	Increase of AKI stage based on KDIGO guidelines
Time	Patients with and without pre-existing AKI stage whose condition worsened after 24 hours of drug administration or within 7 days after the last dose.

Based on the PICOT framework, the causal question is: *among hospitalized patients, what is the causal effect of receiving a suspected nephrotoxic drug (ibuprofen, ketorolac, vancomycin, lisinopril, furosemide, pantoprazole, omeprazole or allopurinol) versus a negative control drug (simethicone, prochlorperazine or lactulose) on the incidence of acute kidney injury occurring after the first day of drug administration?*

2.2. Step 2: Emulation Design

2.2.1. STEP 2A: EMULATION OF TARGET TRIAL PROTOCOL

The cohort includes hospitalized adult patients (age$\geq$ 18 years) with and without pre-existing acute kidney injury, KDIGO stage ≥ 1 and KDIGO stage $= 0$, respectively, before drug administration. The outcome is assessed by the incidence of acute kidney injury, indicated by an increase in KDIGO stage, following the first day of drug administration. Time zero was defined as the time of first in-hospital administration of the suspect or control drug; our cohort includes only patients whose first

3. http://www.nephrotox.com/

exposure was recorded during hospitalization, prior exposure could not be assessed and is noted as a limitation. Included patient characteristics encompass demographics, lab results, concomitant medications (drugs co-administered with the suspect drugs), and comorbidities. Table 2 describes the features included in the final 24 datasets (8 suspect x 3 negative control drugs), where the concomitant drugs are mapped with ATC (Anatomical Therapeutic Chemical) codes.

Table 2: The feature categories in the final datasets

Feature Category	**Number of variables**	**Details**
Demographics	2	Age, gender
Vitals	3	Diastolic BP, systolic BP, weight
Laboratory Tests	8	Glucose, sodium, creatinine, potassium, Blood Urea Nitrogen (BUN), bicarbonate, chloride, anion gap
Concomitant drugs	71 to 472	ATC codes (number of drugs differ)
Comorbidities	17	Charlson comorbidity (Charlson et al., 1987)

For the list of positive and negative control drugs, the included characteristics of the patients are collected only for the period of the drug administered to patients without AKI and until the event for patients with AKI. Data quality processes include removal of EHRs with missing entries in the included features and mapping drug names to ATC codes (Appendix section A.3).

2.2.2. STEP 2B: IDENTIFYING FIT-FOR-PURPOSE DATA SOURCES

Appendix table 5 outlines how the selected data are aligned with the objectives of the study. Summarizing its contents, we conclude that MIMIC-IV database offers comprehensive patient data, including lab tests, medications, comorbidities, KDIGO-defined AKI outcomes, precise drug-administration timing, and other relevant features, which have been evaluated from scientific literature and identified as potentially useful (e.g., as potential confounders). These features allow us to define eligibility windows, specify treatment initiation and duration, identify outcomes using established clinical criteria, and adjust for a pre-specified set of measured confounders.

2.2.3. STRUCTURAL CAUSAL MODEL

To identify potential confounders, colliders, and mediators, we constructed structural causal models (Wang et al., 2025), represented as directed acyclic graphs (DAGs). This procedure follows four distinct steps (Appendix figures 7 and 8): (1) Literature review, (2) Statistical analysis, (3) Literature analysis, (4) Domain expertise. The first step, based on a statistical analysis (Appendix table 4), an initial directed graph is constructed, where all connections are drawn according to the significance of the continuous and categorical variables to the exposure (suspect drug) and the response (AKI). The concomitant drugs that are used as potential confounders are retrieved from Fernández-Llaneza et al. (2024) study identified as potential risk factors for AKI and sources of drug–drug interactions. Next, a supplementary review of the existing scientific literature was conducted to verify the edges of the graph. Finally, in collaboration with a nephrologist, we constructed the final DAGs for each suspect drug.

2.3. Step 3: Assess expected precision and conduct diagnostic evaluations

Eliminating patients with missing entries from our final datasets led to smaller cohorts than initially retrieved from MIMIC-IV, but possible biases from imputation methods were avoided. Appendix table 7 presents the cohort sizes after data preprocessing. Moreover, valid causal inference on observational data requires assumptions. The causal assumptions made for this study are described in the Appendix section E.

We determined the minimal detectable relative risk (MDRR) (Armstrong, 1987), defined as the lowest relative risk identifiable with a significance level (α) of 0.05 and a statistical power of 0.80. The smaller the MDRR is, the more representative the cohort is for detecting small effects. Thus, we adopted the same strict threshold of MDRR ≤ 1.25 as utilized in the Ryan et al. (2013) study. All different cohorts (24) used in our study did not cross this threshold (MDRR ≤ 1.12), as presented in Appendix table 8.

2.4. Step 4: Robustness evaluations

In this step we applied three types of deterministic sensitivity analysis. First, we estimated treatment effects using alternative representations of laboratory measurements, including mean values, the first measurements closest to or on the day of initial drug exposure, and the last measurements prior to AKI occurrence or the final drug dose (Appendix figure 6). Second, we re-estimated effects using two different confounder adjustment sets: one derived from statistical analyses and literature review, and another proposed by a nephrologist for drugs where disagreement existed between the literature-based and clinician-defined confounders. In these analyses, concomitant medications were accounted for using dosing duration (in days) and encoded as binary indicators. Third, we repeated all analyses using normalized laboratory and vital sign measurements and compared them with analyses based on the original (unnormalized) values.

2.5. Step 5: Inferential Analysis

Using machine learning algorithms, CML models estimate treatment effects conditional on patient characteristics (Oprescu et al., 2019). We applied multiple CML approaches (Appendix section D), including meta-learners (S-,T- and X-learners) and double machine learning (DML). Propensity score matching was applied in all datasets to adjust for confounders. For each conditional average treatment effect (CATE) estimate, confidence intervals were computed via bootstrap sampling (n = 100). The CATE was estimated with meta-learners for matched and unmatched samples. The best-performing CML architectures, selected based on confidence intervals' (CI) width and robustness to confounding, were then used to compute heterogeneous treatment effects (HTEs). All CATE estimators (CML models) were implemented using the Econml library (Oprescu et al., 2019).

2.5.1. MATCHING

Propensity Score Matching (PSM) is used to match covariates (Algorithm 3). PSM estimates the probability of receiving treatment using various classifiers, a caliper (coefficient=0.2), and a fixed ratio 4:1. The ML algorithms that are tested: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classifier (SVC), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Multilayer Perceptron (MLP). The propensity score is a balancing score: conditional on it, the distribution of covariates should be similar across treatment groups. Matching

treated and control units with similar propensity scores mimics randomization. Standardized difference in means (SMD) is used for the evaluation of the quality of matching (Equation 2). A moderate SMD cutoff of 0.2 has been selected to indicate the balance of covariates (Austin, 2009).

2.5.2. SELECTION OF ML MODELS FOR CML ARCHITECTURES

A prerequisite to the application of CML methods is to select ML algorithms that best fit the data and the methodologies. So, we tested several ML algorithms (Appendix figure 9), from simple (e.g, Logistic regression) to more complex (e.g., Multilayer Perceptron). Algorithms are selected based on accuracy, recall, precision, and f1 score for classification models and mean squared error (MSE), root mean squared error (RMSE) and R^2 for regression models.

For the S-learner, which uses a single classifier, model selection was based on the average of four classification metrics. The T-learner requires two classifiers (for $T = 0$ and $T = 1$), so performance was computed as the average of both models. The DML-learner requires three regression models; because regression metrics cannot be directly averaged, we defined a combined score S (Equation 1) and used the mean of the three S values. The X-learner uses two classifiers and two regressors; thus, classification performance was assessed as in the T-learner and regression performance as in the DML-learner, with the final score obtained by summing both components.

2.5.3. EVALUATION OF THE RESULTS

The results are evaluated in 2 different directions, the technical evaluation of CML methodologies and the domain evaluation. The CML methodologies are examined based on CI. CATE estimators' CIs' width that exceeded 0.85 are considered non-robust and are unstable estimators (we select this width because it is the largest CI width among significant CATEs whose intervals exclude zero). For clinical validation, HTEs are used in characterizing the heterogeneous treatment effect in different groups of patients based on age, gender, weight, vital signs, and laboratory tests' values. Treatment effects in various patient categories are evaluated. The calculated HTEs are investigated on the groups of patient characteristics (age category, glucose categories etc.) and on HTE categories. The HTEs are separated in 3 main categories: Protective effect where mean HTE $<$ -0.1, no effect where -0.1 $\geq$ mean HTE $\leq$ 0.1 and adverse effect where mean HTE $>$ 0.1. For each patient characteristic the statistical significance is calculated between the different HTE categories with the Kruskal-Wallis test (McKight and Najab, 2010).

3. Results

3.1. Causal graphs

Nephrologist and literature agree on the structure of DAGs for omeprazole, pantoprazole, and vancomycin drugs. However, the structure of DAGs for furosemide, allopurinol, ibuprofen, lisinopril, and ketorolac differs between nephrologist and the existing literature. An interesting disagreement was in allopurinol, where, based on several studies, there is a potential direct connection between the drug and AKI (Appendix figure 20), but the nephrologist disagrees (Appendix figure 21).

In the case of ibuprofen, the literature mentions (Appendix figure 10) that abnormal glucose metabolism occurs at the onset, progression, and prognosis of several kidney diseases like AKI (Wen et al., 2021), but the nephrologist (Appendix figure 11) questions this connection. The same disagreement is met in ketorolac's DAG (Appendix figures 12, 13). In lisinopril, except for the glucose

disagreement, the causal relationship between anti-gout preparations and AKI is also questioned by the nephrologist. In literature (Appendix figure 15), there is evidence that anti-gout medication is directly related to AKI (Rey et al., 2019). In contrast, the nephrologist (Appendix figure 15) supports that this connection can be achieved through the administration of these drugs to comorbidities. Another difference is in the causal association of psycholeptics and AKI, where literature supports that atypical antipsychotic drug use versus non-use was associated with a higher risk of hospitalization with AKI (Hwang et al., 2014). However, the nephrologist did not support this association as psycholeptics are a major class of drugs and there is no strong evidence. In the furosemide DAG except for the anti-gout medication difference, we met a different approach in the causal association of antiepileptics and AKI (Appendix figure 17) where according to the study of Hamed (2017) some patients with epilepsy develop clinical or subclinical kidney dysfunction or injury with long-term use of antiepileptics. This argument is not supported by the nephrologist (Appendix figure 18).

Although graphs may differ between literature and nephrologist, confounders in some cases remain exactly the same (furosemide). Appendix table 9 describes the confounders that are controlled for each drug based on DAGs, both from the literature review and the knowledge of the physician.

3.2. Matching

Across all drug–control comparisons, propensity score matching (PSM) substantially reduced covariate imbalance, with performance varying by preprocessing strategy, propensity score model, and laboratory-test representation. Logistic Regression and Support Vector Classifiers provided the most stable matching across settings, whereas Gradient Boosting and AdaBoost performed better in settings with limited covariate overlap between treatment and control groups. Preprocessing also influenced match quality: *original* covariates preserved larger matched cohorts, *binary* transformations improved balance for scale-sensitive models, and *normalized* features yielded the lowest residual SMD in high-dimensional settings.

When laboratory tests were represented as *mean values* (Appendix table 10), most drugs achieved strong post-matching balance: Ibuprofen (initial SMD 0.18–0.42; matched mean SMD 0.01–0.09; 0–8 covariates with SMD$>$ 0.2), Ketorolac (0.08–0.28; 0.02–0.09; 2–12), Lisinopril (0.14–0.23; 0.01–0.06; 1–6), Allopurinol (0.14–0.22; 0–0.04; 0–8), and Vancomycin (0.14–0.36; 0–0.08; 0–3). In contrast, drugs such as Furosemide (0.20–0.34; 0–0.17; 4–15), Pantoprazole (0.11–0.41; 0–0.22; 1–18), and Omeprazole (0.09–0.35; 0–0.21; 1–21) showed higher residual imbalance. Using first/last laboratory values (Appendix table 11) instead of means consistently worsened match quality, increasing mean post-matching SMD and the number of covariates with SMD$>$ 0.2, indicating that mean-value aggregation provides a more stable confounder representation.

3.3. Machine-learning model selection to support CML

Across all learning architectures evaluated, S-learner (Appendix table 12), T-learner (Appendix table 13), X-learner (Appendix table 14), and DML (Appendix table 15), the Random Forest model consistently emerged as the top performer, particularly in the S-learner where it delivered near-perfect predictive accuracy. Although model performance varied more in T-learner and X-learner, Random Forest and XGBoost remained among the most reliable algorithms, with additional contributions from Decision Trees, SGD, and Support Vector methods depending on the control drug and drug-specific dataset. Notably, X-learners and DML approaches introduced greater model diversity and flexibility through combined classification-regression strategies and multi-stage estimation

pipelines. In classification models, the X-learner's performance was generally lower than in the S-learner and T-learner settings. The regression model performs better in DML than the X-learner.

3.4. Conditional Average Treatment Effect

The results are classified as follows: Original (O) refers to the initial datasets where concomitant drugs include the number of days administered; Binary (B) refers to datasets where concomitant drugs are represented as binary variables (0/1); and Normalized (N) refers to datasets where concomitant drugs are binary and continuous variables (laboratory tests, vital signs) are normalized. *Clinicians* indicates confounders selected using the nephrologist-designed DAG, while *Literature* indicates confounders selected using the literature-based DAG. Laboratory tests are represented either as mean values (Mean) or as first and last measurements (t0–t1). Only CATE results from unmatched samples were analyzed, as matched analyses did not yield valid results for all suspected drugs (Appendix section I).

Following the initial CATE evaluation, results with CI < 0.85 were considered valid. Figure 2 shows the distribution of valid CATE pipelines. Ibuprofen (86/144 CATEs) had the most valid results, particularly with simethicone. Nephrologist DAGs outperformed literature DAGs. Laboratory result representation had minimal impact, though mean values produced slightly more valid CATEs than first/last values. The S-learner generated the most robust ATEs.

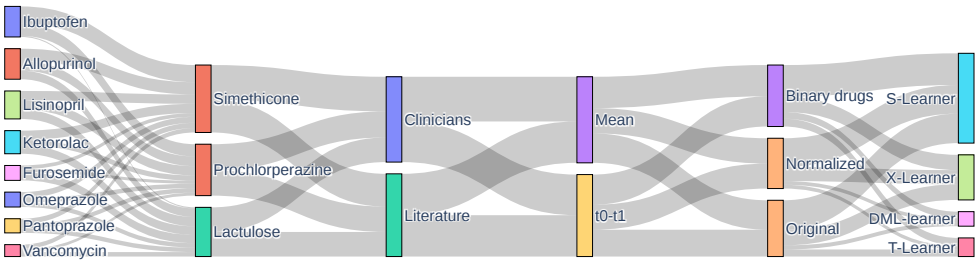

Figure 2: Pipelines of work based on the number of valid CATE results (CI range < 85)

Valid CATEs are evaluated by their statistical significance. A CATE is considered statistically significant when the CIs range does not include 0. Figure 3 illustrates only 4 out of 8 suspect drugs (lisinopril, ibuprofen, furosemide, ketorolac), including significant CATEs. Lisinopril shows the most significant results in combination with negative drugs prochlorperazine and simethicone but lactulose is the negative drug that includes the most significant results in total. The most significant results are shown in DAGs designed by literature in combination with normalized values in laboratory tests and vital signs where laboratory tests are represented as first and last values. X-learner produces the most significant CATEs where most of them demonstrate negative effect to AKI.

Ranking suspect drugs by the total average CATE of all CML methods reveals negative effects for 3 out of 8 drugs (allopurinol, lisinopril, ketorolac) with CATE values ranging from -0.04 to -0.17. Selecting the CML architecture for each drug that reflects the possibility of a positive effect (DAKI), as described in the relevant literature, resulted in a drug ranking that aligned better with

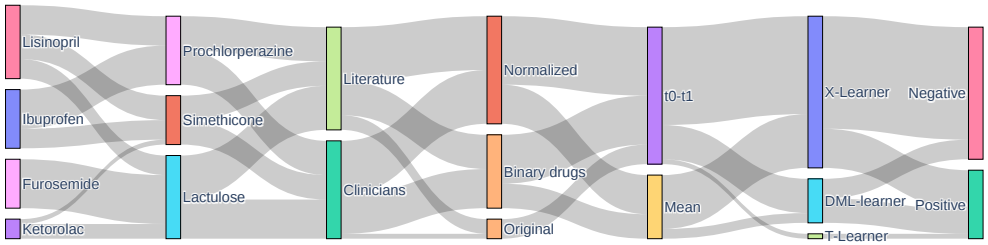

Figure 3: Significant CATEs (CI excludes 0) Estimates (with 95% CI)

the effect ranking in Fernández-Llaneza et al. (2024). Ibuprofen and furosemide CATEs have the highest standard deviations. Figure 31 describes in detail the average value of CML methods for each suspect drug.

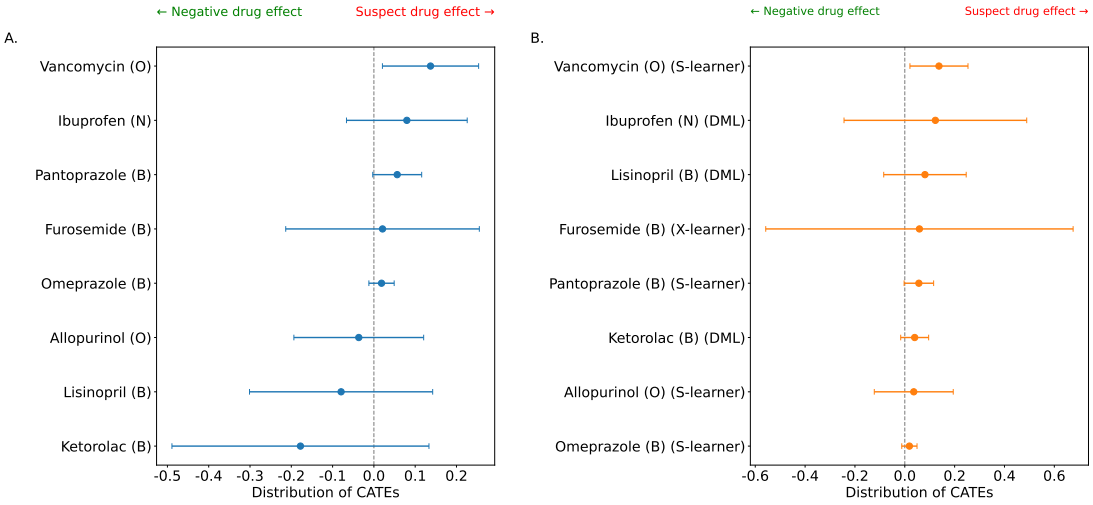

Figure 4: Ranking of drugs A. Mean ± SD of total CATE estimates per suspect drug. B. Mean ± SD of positive CATEs across CML methods per suspect drug.

3.5. Heterogeneous Treatment Effect

CML methodologies were selected for HTE estimation based on ATE CI width (<0.85), significance (CIs not containing 0), and consistency with the literature (Figure 5). For example, although Ketorolac mostly yielded negative CATEs, the X-learner with mean lab values, lactose as the negative drug, and the literature DAG produced a positive CATE of 0.2 (-0.02, 0.4), compared to the DML learner, which gave a smaller CATE of 0.07 with wider CI (-0.3, 0.5). Due to methodological diversity, HTE results are discussed as separate case studies (Appendix Table 16).

After applying Kruskal-Wallis statistical testing for the 3 categories of HTE values (protective effect, no effect and adverse effect), the characteristics of patients with suspect DAKI are described per drug (For allopurinol, no statistically significant patient characteristics were identified). Appendix section K describes the analysis of the characteristics of patients based on HTE value in more detail.

Ibuprofen (150 AKI patients): Higher HTEs were observed in males (mean 0.4, SD 0.1) compared to females (0.3, SD 0.1), and in adults aged 40–59 (0.4, SD 0.05). Overweight patients showed slightly higher effects (0.34, SD 0.1) than obese patients (0.31, SD 0.1). Elevated HTEs were associated with increased glucose (0.4, SD 0.1), creatinine $\geq$ 2.0 mg/dL (0.4, SD 0.05), BUN $>$ 20 mg/dL (0.4, SD 0.05), bicarbonate (0.34, SD 0.1), potassium (0.32, SD 0.1), low chloride (0.34, SD 0.1), and lower anion gap (0.3, SD 0.1). Both systolic and diastolic blood pressures were slightly higher (0.3, SD 0.1).

Ketorolac (399 AKI patients): Males presented higher HTEs (0.3, SD 0.1) compared to females (0.2, SD 0.2), and adults $\geq$ 80 years had the highest HTE (0.3, SD 0.2). Increased glucose (0.3, SD 0.2), creatinine $\geq$ 2.0 mg/dL (0.3, SD 0.1), BUN $>$ 20 mg/dL (0.35, SD 0.1), bicarbonate (0.3, SD 0.2), and potassium (0.3, SD 0.2) corresponded to higher effects. Reduced chloride (0.3, SD 0.2), anion gap (0.2, SD 0.2), and sodium (0.3, SD 0.15) were also observed. Blood pressure was elevated in the adverse group (0.2, SD 0.2).

Vancomycin (2453 AKI patients): Higher HTEs were found in young adults (18–39; 0.4, SD 0.5), females (0.3, SD 0.5), and overweight/obese patients (0.3, SD 0.5). Most laboratory markers were lower in the adverse group, including creatinine (0.3, SD 0.5), glucose (0.3, SD 0.5), potassium (0.4, SD 0.5), BUN (0.6, SD 0.5), bicarbonate (0.3, SD 0.5), and anion gap (0.3, SD 0.5). Chloride was the only increased marker (0.3, SD 0.5). Diastolic blood pressure was elevated (0.3, SD 0.4).

Lisinopril (1420 AKI patients): Higher HTEs occurred in average-weight patients (60–79 kg; 0.5, SD 0.3), adults aged 40–59 (0.6, SD 0.2), and females (0.5, SD 0.4). Elevated creatinine $\geq$ 2.0 mg/dL (0.5, SD 0.3), anion gap (0.5, SD 0.3), and glucose (0.5, SD 0.3) were associated with higher HTEs. Decreased bicarbonate (0.5, SD 0.3), chloride (0.5, SD 0.3), BUN (0.6, SD 0.2), and sodium (0.5, SD 0.35) were also noted. Diastolic blood pressure was high (0.5, SD 0.35).

Furosemide (2669 AKI patients): Adults aged 40–59 showed the highest HTEs (0.6, SD 0.2). Significant differences were found in bicarbonate (0.5, SD 0.35), BUN (0.5, SD 0.3), and sodium (0.5, SD 0.35), all of which increased in the adverse group. Systolic blood pressure was lower (0.5, SD 0.35), while diastolic pressure was elevated (0.5, SD 0.35).

Pantoprazole (2328 AKI patients): Patients in the adverse group were young adults (18–59; 0.01, SD 0.2), female (0.001, SD 0.1), and exhibited lower bicarbonate levels (0.001, SD 0.1).

Omeprazole (1560 AKI patients): Similar to pantoprazole, adverse-effect patients were young adults (18–59; 0.02, SD 0.15), mostly female (0.01, SD 0.1). Chloride was increased (0.005, SD 0.1). Lower bicarbonate (0.01, SD 0.1), BUN (0.1, SD 0.3), glucose (0.01, SD 0.1), and creatinine (0.003, SD 0.1) levels were observed. Systolic blood pressure was lower (0.002, SD 0.1).

4. Discussion

Our framework provides a structured approach for extracting reliable pharmacovigilance signals from EHRs, which offer richer clinical information than traditional sources that often yield biased, non-causal associations. Using the PRINCIPLED workflow (Desai et al., 2024) with CML, we can reduce bias. Although EHR data is not publicly accessible, the protocol-driven design and detailed

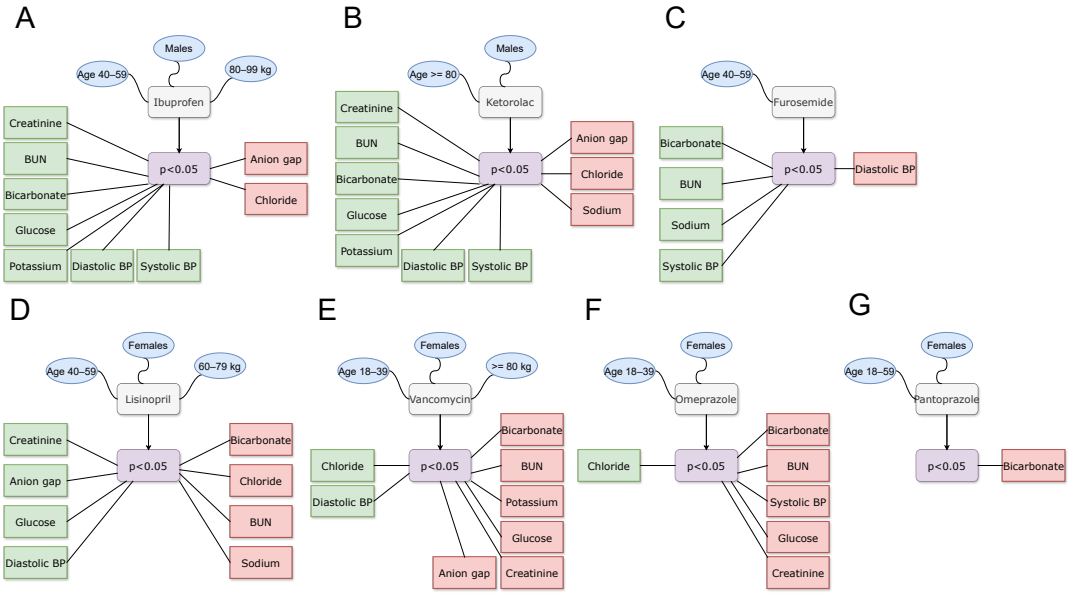

Figure 5: Analysis of heterogeneous treatment effect (HTE) of suspect drugs with statistically significant (Kruskal-Wallis test) patient characteristics where HTE > 0.1 (AKI adverse effect). Red indicates decreased and green increased values. A. Ibuprofen B. Ketorolac C. Furosemide D. Lisinopril E. Vancomycin F. Omeprazole G. Pantoprazole

variable definitions support reproducibility. Additionally, our use of comprehensive evaluation metrics on both ML and CML models helps limit false-positive safety signals.

4.1. Addressing confounding and model robustness in CML estimation

Minimizing confounding bias is essential for reliable evidence, but real-world data contain unobserved factors that cannot be fully addressed by structural causal models or existing de-confounding methods (Kuzmanovic et al., 2021; Bica et al., 2020). To reduce residual bias, we used three negative-control drugs. Matching results showed variability on SMDs and the number of matched patients (less than 200) across controls and ML models (Appendix tables 11, 10), highlighting the impact of the selection of controls.

We further addressed confounding by constructing two DAGs per suspect drug, based on literature and nephrologist knowledge. Agreement was reached for pantoprazole, omeprazole, furosemide, and vancomycin, while ibuprofen, ketorolac, lisinopril, and allopurinol showed structural differences. Nonetheless, CATE estimates were generally consistent between the two DAGs, with divergences appearing mainly in the width of confidence intervals when effects were statistically significant. This may indicate that differences between DAG specifications are minor, as the additional confounders represent weak causal links.

As no gold-standard CML method exists for real-world data, we evaluated four learners across all DAKI cases. After filtering invalid estimates (CI ≥ 0.85), the S-learner retained the most valid not significant CATEs (~ 0), indicating robustness. The T-learner performed worst, with wide CIs

and few valid estimates likely due to bias from separate outcome models. The X-learner produced the second-highest number of valid estimates and produced significant CATEs for furosemide (-0.4 to 0.5) and lisinopril (-0.2) with lactulose. DML also gave significant results, particularly for ibuprofen (0.3–0.5) and lisinopril (0.3) with prochlorperazine. Overall, complex estimators, X-learner and DML, captured the only significant effects, S-learner robust but uninformative, and T-learner was the least reliable.

4.2. CML can support pharmacovigilance

We compared our CATE-based rankings with those reported in Fernández-Llaneza et al. (2024), and evaluated HTE patterns across patient subgroups against existing findings in the literature. For ketorolac, the CML models used in the HTE analysis were chosen based on the available domain knowledge about DAKI.

Across suspect drugs, the closest agreement with the literature occurred when using literature-based DAGs, first–last laboratory values, and the CML method with the most statistically significant CATEs. Ibuprofen consistently showed a strong AKI effect in both studies, while vancomycin and pantoprazole showed positive but non-significant effects. Lisinopril produced mixed but sometimes significant positive CATEs. Furosemide exhibited opposite significant effects depending on laboratory representation, differing from Fernández-Llaneza et al. (2024), possibly due to higher AKI incidence in our sample. Omeprazole and allopurinol showed small, non-significant effects consistent with the uncertain evidence in prior work.

HTE patterns were partly supported by known risk factors. For gender, ibuprofen and ketorolac exhibited higher risk for men, consistent with Faguer et al. (2024). Age effects aligned with the literature for ketorolac (Klomjit and Ungprasert, 2022) and furosemide (Wu et al., 2014), although ibuprofen and vancomycin showed inconsistencies likely due to small subgroup sizes or sample-specific bias. Lisinopril diverged from expectations (Chen et al., 2021), while omeprazole and pantoprazole displayed too few high-HTE cases for interpretation.

Weight-related HTEs patterns were inconsistent with the obesity-related AKI literature for vancomycin (Rutter et al., 2019; Danziger et al., 2016; Shi et al., 2020), whereas ibuprofen, lisinopril, and furosemide showed significant associations likely influenced by high comorbidity burden. Glucose results were inconsistent with the mechanisms outlined in Wen et al. (2021), suggesting limited relevance of glucose to DAKI in our cohort.

Electrolyte patterns largely reflected known drug mechanisms. Elevated creatinine for ibuprofen, ketorolac, and lisinopril aligned with Zhang et al. (2017); Thorp et al. (2005). Potassium abnormalities were consistent with Man et al. (2022) for ibuprofen, and partially for ketorolac. BUN findings aligned with NSAID nephrotoxicity (Whelton, 1999) and loop-diuretic physiology Nunez et al. (2012), though vancomycin and lisinopril diverged from expected elevations Wood et al. (1986); Ahmed (2002). Bicarbonate results matched literature for ibuprofen (Man et al., 2022), furosemide (Emmett, 2020), and vancomycin (Mori et al., 2018). Chloride and anion-gap differences have not been previously reported. Blood-pressure effects were consistent with NSAID physiology (Warner and Mitchell, 2008) and furosemide's known hemodynamic influence (Osmanovic et al., 2017).

Finally, because causal effects reflect the full causal graph, including mediators, comorbidities, and concomitant drugs, results may vary across datasets. Thus, while CML enables patient-specific inference and subgroup insight, generalizing these effects requires datasets with similar distributions and causal structures.

4.3. Limitations and Future work

The main limitation is the reliance on a single database (MIMIC-IV), which restricts generalizability. The dataset also lacks socioeconomic variables. In preprocessing, text-matching approaches for ATC mapping (85% threshold) may have led to drug mapping errors, and drug dosage information for suspect and control drugs was not available.

Future extensions could incorporate causal discovery methods (Vowels et al., 2022) and causal checking approaches to identify missing relationships and to test whether conditional independence relations implied by manually constructed DAGs are supported by the data. Advanced deconfounding techniques (Louizos et al., 2017; Witty et al., 2020; Kuzmanovic et al., 2021; Bica et al., 2020) to further reduce hidden bias, and more powerful CML estimators such as GANITE (Yoon et al., 2018), Super Learning (Luedtke and van der Laan, 2016), or Generalized Random Forests (Athey et al., 2019).

References

Ali Ahmed. Use of angiotensin-converting enzyme inhibitors in patients with heart failure and renal insufficiency: how concerned should we be by the rise in serum creatinine?, 2002.

Ben Armstrong. A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. *American journal of epidemiology*, 126(2):356–358, 1987.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International conference on machine learning*, pages 884–895. PMLR, 2020.

Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

Jui-Yi Chen, I-Jung Tsai, Heng-Chih Pan, Hung-Wei Liao, Javier A Neyra, Vin-Cent Wu, and Jeff S Chueh. The impact of angiotensin-converting enzyme inhibitors or angiotensin ii receptor blockers on clinical outcomes of acute kidney disease patients: a systematic review and meta-analysis. *Frontiers in Pharmacology*, 12:665250, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Salvatore Crisafulli, Zakir Khan, Yusuf Karatas, Marco Tuccori, and Gianluca Trifirò. An overview of methodological flaws of real-world studies investigating drug safety in the post-marketing setting. *Expert Opinion on Drug Safety*, 22(5):373–380, 2023.

John Danziger, Ken P Chen, Joon Lee, Mengling Feng, Roger G Mark, Leo Anthony Celi, and Kenneth J Mukamal. Obesity, acute kidney injury, and mortality in critical illness. *Critical care medicine*, 44(2):328–334, 2016.

Rishi J Desai, Shirley V Wang, Sushama Kattinakere Sreedhara, Luke Zabotka, Farzin Khosrow-Khavar, Jennifer C Nelson, Xu Shi, Sengwee Toh, Richard Wyss, Elisabetta Patorno, et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (principled): considerations from the fda sentinel innovation center. *bmj*, 384, 2024.

Matthieu Doutreligne, Tristan Struja, Judith Abecassis, Claire Morgand, Leo Anthony Celi, and Gaël Varoquaux. Step-by-step causal analysis of ehrs to ground decision-making. *PLOS Digital Health*, 4(2):e0000721, 2025.

Michael Emmett. Metabolic alkalosis: a brief pathophysiologic review. *Clinical journal of the American Society of Nephrology*, 15(12):1848–1856, 2020.

Stanislas Faguer, Alexis Piedrafita, Ana Belen Sanz, Justyna Siwy, Ioanna K Mina, Melinda Alves, Paul Bousquet, Bertrand Marcheix, Audrey Casemayou, Julie Klein, et al. Performances of acute kidney injury biomarkers vary according to sex. *Clinical Kidney Journal*, 17(5):sfae091, 2024.

Daniel Fernández-Llaneza, Romy MP Vos, Joris E Lieverse, Helen R Gosselt, Sandra L Kane-Gill, Teun van Gelder, and Joanna E Klopotowska. An integrated approach for representing knowledge on the potential of drugs to cause acute kidney injury. *Drug Safety*, pages 1–16, 2024.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Gordon H Guyatt and Drummond Rennie. Users' guides to the medical literature. *Jama*, 270(17): 2096–2097, 1993.

Sherifa Ahmed Hamed. The effect of antiepileptic drugs on the kidney function and structure. *Expert Review of Clinical Pharmacology*, 10(9):993–1006, 2017.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Y Joseph Hwang, Stephanie N Dixon, Jeffrey P Reiss, Ron Wald, Chirag R Parikh, Sonja Gandhi, Salimah Z Shariff, Neesh Pannu, Danielle M Nash, Faisal Rehman, et al. Atypical antipsychotic drugs and the risk for acute kidney injury and other adverse outcomes in older adults: a population-based cohort study. *Annals of internal medicine*, 161(4):242–248, 2014.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55, 2020.

Iman Karimzadeh, Erin F Barreto, John A Kellum, Linda Awdishu, Patrick T Murray, Marlies Ostermann, Azra Bihorac, Ravindra L Mehta, Stuart L Goldstein, Kianoush B Kashani, et al. Moving toward a contemporary classification of drug-induced kidney disease. *Critical Care*, 27 (1):435, 2023.

John A Kellum, Norbert Lameire, Peter Aspelin, Rashad S Barsoum, Emmanuel A Burdmann, Stuart L Goldstein, Charles A Herzog, Michael Joannidis, Andreas Kribben, Andrew S Levey, et al. Kidney disease: improving global outcomes (kdigo) acute kidney injury work group. kdigo clinical practice guideline for acute kidney injury. *Kidney international supplements*, 2(1):1–138, 2012.

Nattawat Klomjit and Patompong Ungprasert. Acute kidney injury associated with non-steroidal anti-inflammatory drugs. *European journal of internal medicine*, 101:21–28, 2022.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Milan Kuzmanovic, Tobias Hatt, and Stefan Feuerriegel. Deconfounding temporal autoencoder: estimating treatment effects over time using noisy proxies. In *Machine Learning for Health*, pages 143–155. PMLR, 2021.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

Alexander R Luedtke and Mark J van der Laan. Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics*, 12(1):305–332, 2016.

Anca M Man, Arianna Piffer, Giacomo D Simonetti, Martin Scoglio, Pietro B Fare, Sebastiano AG Lava, Mario G Bianchetti, and Gregorio P Milani. Ibuprofen-associated hypokalemia and metabolic acidosis: systematic literature review. *Annals of Pharmacotherapy*, 56(11):1250–1257, 2022.

Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

Ravindra L Mehta, Linda Awdishu, Andrew Davenport, Patrick T Murray, Etienne Macedo, Jorge Cerda, Raj Chakaravarthi, Arthur L Holden, and Stuart L Goldstein. Phenotype standardization for drug-induced kidney disease. *Kidney international*, 88(2):226–234, 2015.

Takahiro Mizuguchi and Shoichi Sawamura. Machine learning-based causal models for predicting the response of individual patients to dexamethasone treatment as prophylactic antiemetic. *Scientific Reports*, 13(1):7549, 2023. doi: 10.1038/s41598-023-34505-0. URL https://doi.org/10.1038/s41598-023-34505-0.

Nobuaki Mori, Yoshio Kamimura, Yuki Kimura, Shoko Hirose, Yasuko Aoki, and Seiji Bito. Comparative analysis of lactic acidosis induced by linezolid and vancomycin therapy using cohort and case–control studies of incidence and associated risk factors. *European Journal of Clinical Pharmacology*, 74:405–411, 2018.

Jerzy Neyman and Elizabeth L Scott. Asymptotically optimal tests of composite hypotheses for randomized experiments with noncontrolled predictor variables. *Journal of the American Statistical Association*, 60(311):699–721, 1965.

Julio Nunez, Eduardo Nunez, Gema Minana, Vicent Bodí, Gregg C Fonarow, Vicente Bertomeu-González, Patricia Palau, Pilar Merlos, Silvia Ventura, Francisco J Chorro, et al. Differential mortality association of loop diuretic dosage according to blood urea nitrogen and carbohydrate antigen 125 following a hospitalization for acute heart failure. *European journal of heart failure*, 14(9):974–984, 2012.

Miruna Oprescu, Vasilis Syrgkanis, Keith Battocchi, Maggie Hei, and Greg Lewis. Econml: A machine learning library for estimating heterogeneous treatment effects. In *33rd Conference on Neural Information Processing Systems*, page 6, 2019.

Elvedin Osmanovic, Esed Omerkic, Semir Imamovic, Mirza Mukinovic, Halid Mahmutbegovic, and Mersiha Cerkezovic. The effect of furosemide on arterial blood pressure, blood glucose levels and incidence of heart arrhythmias. *American Journal of Internal Medicine*, 5(3):41–45, 2017.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

Amayelle Rey, Benjamin Batteux, Solène M Laville, Justine Marienne, Kamel Masmoudi, Valérie Gras-Champel, and Sophie Liabeuf. Acute kidney injury associated with febuxostat and allopurinol: a post-marketing study. *Arthritis Research & Therapy*, 21:1–9, 2019.

W Cliff Rutter, Ronald G Hall 2nd, and David S Burgess. Impact of total body weight on rate of acute kidney injury in patients treated with piperacillin–tazobactam and vancomycin. *American Journal of Health-System Pharmacy*, 76(16):1211–1217, 2019.

Patrick B Ryan, Martijn J Schuemie, Emily Welebob, Jon Duke, Sarah Valentine, and Abraham G Hartzema. Defining a reference set to support methodological research in drug safety. *Drug safety*, 36:33–47, 2013.

Ning Shi, Kang Liu, Yuanming Fan, Lulu Yang, Song Zhang, Xu Li, Hanzhang Wu, Meiyuan Li, Huijuan Mao, Xueqiang Xu, et al. The association between obesity and risk of acute kidney injury after cardiac surgery. *Frontiers in endocrinology*, 11:534294, 2020.

Paweena Susantitaphong, Dinna N Cruz, Jorge Cerda, Maher Abulfaraj, Fahad Alqahtani, Ioannis Koulouridis, and Bertrand L Jaber. World incidence of aki: a meta-analysis. *Clinical journal of the American Society of Nephrology*, 8(9):1482–1493, 2013.

ML Thorp, DG Ditmer, MK Nash, R Wise, PL Jaderholm, JD Smith, and W Chan. A study of the prevalence of significant increases in serum creatinine following angiotension-converting enzyme inhibitor administration. *Journal of human hypertension*, 19(5):389–392, 2005.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Sushrut S Waikar, Kathleen D Liu, and Glenn M Chertow. Diagnosis, epidemiology and outcomes of acute kidney injury. *Clinical Journal of the American Society of Nephrology*, 3(3):844–861, 2008.

Linbo Wang, Thomas Richardson, and James Robins. Causal inference: A tale of three frameworks. *arXiv preprint arXiv:2511.21516*, 2025.

Shirley V Wang, Sushama Kattinakere Sreedhara, and Sebastian Schneeweiss. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nature communications*, 13(1):5126, 2022.

Yu Wang, Jing Ma, Shuang Ma, Jiaqi Wang, and Jingsong Li. Causal evaluation of post-marketing drugs for drug-induced liver injury from electronic health records. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

Timothy D Warner and Jane A Mitchell. Cox-2 selectivity alone does not define the cardiovascular risks associated with non-steroidal anti-inflammatory drugs. *The Lancet*, 371(9608):270–273, 2008.

Lu Wen, Ying Li, Siyao Li, Xiaoru Hu, Qingqing Wei, and Zheng Dong. Glucose metabolism in acute kidney injury and kidney repair. *Frontiers in medicine*, 8:744122, 2021.

Andrew Whelton. Nephrotoxicity of nonsteroidal anti-inflammatory drugs: physiologic foundations and clinical implications. *The American journal of medicine*, 106(5):13S–24S, 1999.

WHO et al. Anatomical therapeutic chemical (atc) classification, 2021.

Sam Witty, Kenta Takatsu, David Jensen, and Vikash Mansinghka. Causal inference using gaussian processes with structured latent confounders. In *International Conference on Machine Learning*, pages 10313–10323. PMLR, 2020.

CA Wood, SJ Kohlhepp, PW Kohnen, DC Houghton, and DN Gilbert. Vancomycin enhancement of experimental tobramycin nephrotoxicity. *Antimicrobial agents and chemotherapy*, 30(1):20–24, 1986.

Xiaojing Wu, Wen Zhang, Hong Ren, Xiaonong Chen, Jingyuan Xie, and Nan Chen. Diuretics associated acute kidney injury: clinical and pathological analysis. *Renal failure*, 36(7):1051–1055, 2014.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

Xiaoxi Zeng, Gearoid M McMahon, Steven M Brunelli, David W Bates, and Sushrut S Waikar. Incidence, outcomes, and comparisons across definitions of aki in hospitalized individuals. *Clinical Journal of the American Society of Nephrology*, 9(1):12–20, 2014.

Jianqiu Zhang, Erich Kummerfield, Gretchen Hultman, Paul E Drawz, Terrence J Adam, Gyorgy Simon, and Genevieve B Melton. Application of causal discovery algorithms in studying the nephrotoxicity of remdesivir using longitudinal data from the ehr. In *AMIA Annual Symposium Proceedings*, volume 2022, page 1227, 2023.

Xinyu Zhang, Peter T Donnan, Samira Bell, and Bruce Guthrie. Non-steroidal anti-inflammatory drug induced acute kidney injury in the community dwelling general population and people with chronic kidney disease: systematic review and meta-analysis. *BMC nephrology*, 18:1–12, 2017.

Appendix A. PRINCIPLED

A.1. Specification of PICO(T) target trial protocol

Table 3: PICOT framework for drug-induced acute kidney injury (DAKI) in electronic health records (MIMIC-IV) based on the format that is presented from Doutreligne et al. (2025) study.

PICOT	**Description**	**Notation**	**Application in DAKI prediction**
Population	What is the target population of interest?	$X \sim P(X)$, the covariate distribution	Hospitalized patients with AKI stage ≥ 1 based to KDIGO guidelines
Intervention	What is the treatment?	$A \sim P(A = 1) = pA$, the probability to be ADR	Treatment with Ibuprofen, Ketorolac, Vancomycin, Lisinopril, Furosemide, Pantoprazole, Omeprazole, and Allopurinol drugs
Control	What is the clinically relevant comparator?	$1 - A \sim 1 - pA$	Treatment with Simethicone, Prochlorperazine, and Lactulose which are known from the literature for no correlation or causation with AKI outcome
Outcome	What are the outcomes to compare?	$Y(1), Y(0) \sim P(Y(1), Y(0))$, the potential outcomes distribution	AKI stage ≥ 1
Time	Is the start of follow-up aligned with the intervention assignment?	N/A	Patients with and without pre-existing AKI stage whose condition worsened after 24 hours of drug administration or within 7 days after the last dose.

A.2. Emulation Design

Table 4: List of tables with the statistical analysis of patient characteristics in correlation with treatment (suspect drug) and outcome (acute kidney injury) - Link to github.

Table 5: Determining fit-for-purpose data sources (step 2b of the process guide for inferential studies using healthcare data from routine clinical practice, Desai et al. (2024))

Questions	Study's aswers
Q1. Can eligibility criteria be emulated with sufficient accuracy?	The MIMIC-IV database contains all essential patient medical record information needed to meet the eligibility criteria of this study, such as KDIGO guidelines.
Q2. Is outcome of interest measured with sufficient quality?	Acute kidney injury is detected in each patient through KDIGO guidelines that are included in a separate table in the MIMIC-IV database.
Q3. Is treatment measured with sufficient quality?	The suspect drugs (treatment) are sufficiently measured as administered at least 24h before the outcome.
Q4. Are key confounders recorded?	The respective Directed Acyclic Graphs (DAGs) are designed for each drug event combination, based on existing literature and clinical insights provided by a medical doctor with significant real-world experience.

Table 6: Included positive control drugs characteristics in MIMIC IV before preprocessing

Drug name	**# of AKI patients in MIMIC-IV**	**Drug effect**	**Drug class**
Ibuprofen	1361	Very strong	NSAID
Ketorolac	1363	Very strong	NSAID
Vancomycin	11149	Strong	Antibacterial
Lisinopril	4521	Strong	RAAS-acting agent
Furosemide	11065	Moderate	Diuretic
Pantoprazole	8336	Moderate	PPI
Omeprazole	5255	Moderate	PPI
Allopurinol	1824	Limited	Xanthine oxidase inhibitor

A.3. Data preprocessing pipeline

The data preprocessing procedure is outlined in detail in the steps below:

1. Identification of patient characteristics in the EHR with the least missing values, ensuring that a sufficient number of patients can be included in the analysis. Table 2 presents the final input features that are selected.

2. In laboratory tests, values are collected for the period after drug exposure till the time of AKI incidence or till the end of drug exposure. In the final datasets, the value on the day of drug

Table 7: The number of AKI patients after preprocessing in the final datasets

Drug name	Simethicone	Prochlorperazine	Lactulose
Ibuprofen	152	150	151
Ketorolac	408	376	399
Vancomycin	2592	2617	2453
Lisinopril	1420	1420	1420
Furosemide	2753	2846	2669
Pantoprazole	2328	2321	2323
Omeprazole	1551	1559	1560
Allopurinol	734	734	734

exposure is included as well as the last value before AKI incidence for each patient, and the mean value of drug administration till the time of AKI incidence.

3. Vital signs are collected only for the period of drug administration till the time of AKI incidence, and the mean value is calculated for each patient.

4. Concomitant medications—drugs co-administered with each suspect or control drug—vary across datasets. Because drug names in MIMIC-IV are free text, text-mining preprocessing is required. Algorithm 1 outlines the procedure used to standardize and identify relevant drugs, which was manually validated for each dataset.

Algorithm 1 Unify Drug Names

Require: List of drug names D
Ensure: Unified set of drug names U
Remove drug names with length ≤ 3 from D
Initialize an empty set U to store unified drug names
Sort D in ascending order of length
for each drug d in D **do**
    Check if d is a substring of any existing drug in U
    if d is not a substring of any element in U **then**
        Add d to U
    end if
end for
return U

5. Free-text drug names are mapped to the Anatomical Therapeutic Chemical (ATC) classification (WHO et al., 2021). Approximate matching to official ATC names is performed using an 85% string-similarity threshold, selected after data-quality evaluation. The procedure is implemented in Algorithm 2 and was manually validated for each drug dataset.

6. For each concomitant drug administered to a patient, the number of days of exposure is recorded until the occurrence of an AKI event.

7. Extraction of common patient IDs between the tables of patients' characteristics and adding all the information in one final dataset.
8. A similar procedure is followed in the datasets for patients who took the negative or positive control drugs, and did not lead to an AKI outcome.
9. Align the datasets with common features (concomitant drugs) between the datasets of AKI and non-AKI patients for each one of the negative and positive control drugs.
10. Concatenate datasets of AKI and non-AKI patients to one dataset for each drug.

Algorithm 2 Replace Drug Name with ATC Code

Require: Drug name d, Dictionary of ATC codes A
Ensure: Corresponding ATC code or original drug name
Find the most similar match m for d in the keys of A using Unify Drug Names function
Compute the similarity score s for the match
if $s \geq 85$ **then**
  return Corresponding ATC code $A[m]$
else
  return Original drug name d
end if

After preparing one dataset for each positive and negative drug, we generated the final datasets by combining every positive–negative drug pair (Appendix table 7). To avoid introducing bias through imputation, patients with missing values were omitted. The process involved three steps: (1) removing patients who received both the positive and negative drug simultaneously, (2) excluding concomitant-drug variables that were not shared between the paired datasets, and (3) concatenating the resulting datasets to form the final analysis sets.

Figure 6 illustrates the information included in the final datasets.

A.4. Precision Diagnostics - Minimal detectable relative risk

Table 8: Minimal detectable relative risk (MDRR) in different suspect and control drug combinations

Drug name	Simethicone	Prochlorperazine	Lactulose
Ibuprofen	1.12	1.10	1.10
Ketorolac	1.10	1.10	1.10
Vancomycin	1.07	1.07	1.06
Lisinopril	1.07	1.07	1.06
Furosemide	1.06	1.08	1.06
Pantoprazole	1.06	1.05	1.04
Omeprazole	1.07	1.04	1.03
Allopurinol	1.10	1.11	1.11

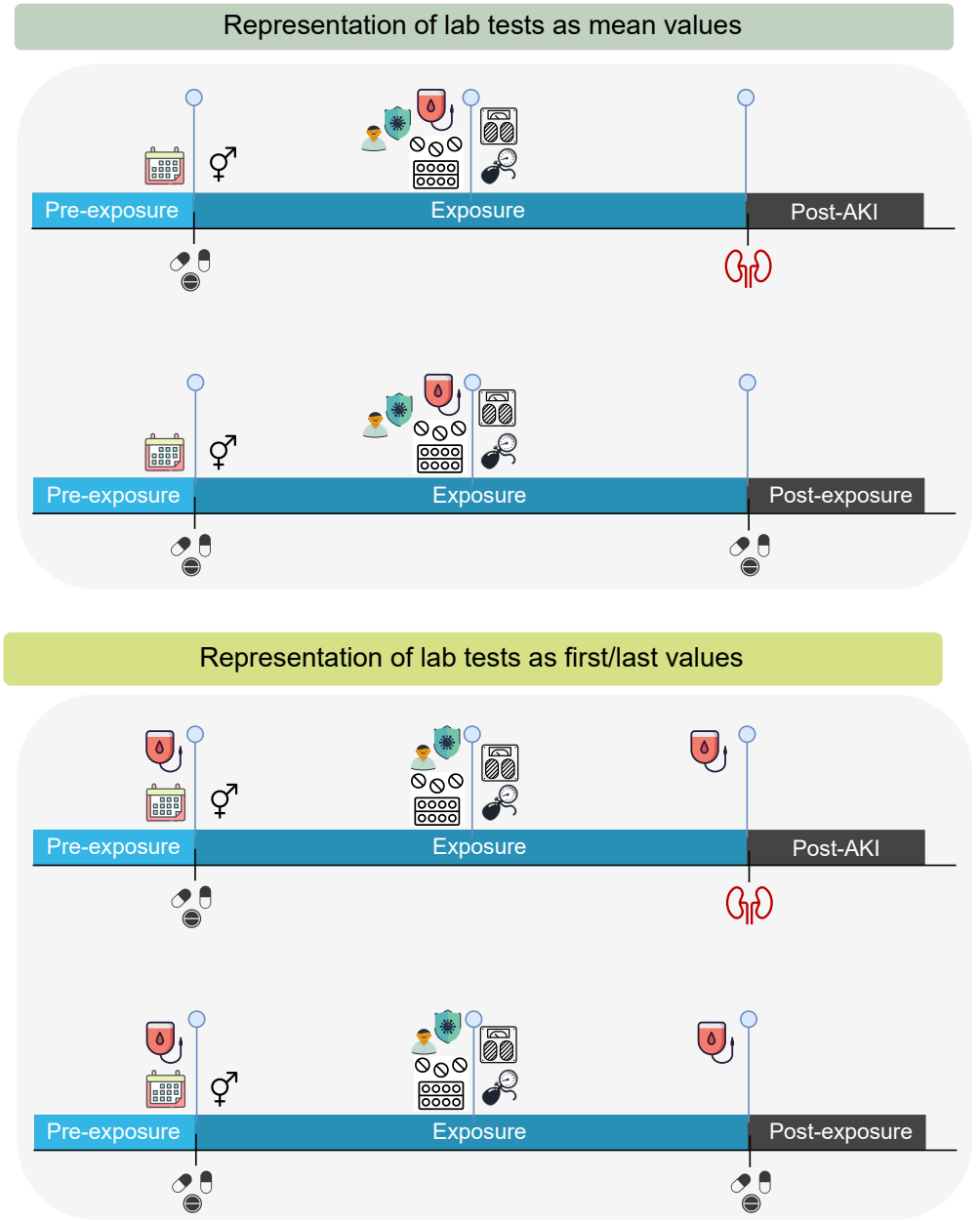

Figure 6: The time representation of patients characteristics in the final datasets

Appendix B. Causal graphs design pipeline

Based on the described workflow in Appendix figure 7, a multi-stage process is employed to construct causal graphs for drug-induced acute kidney injury (DAKI). The process began with a lit-

erature review, in which existing research on DAKI was systematically examined to identify relevant variables and previously reported associations between drug exposures and AKI outcomes. Next, statistical analyses were conducted to empirically assess variable relationships between suspect drugs and AKI outcome using significance tests (e.g., Welch's t-test and Pearson's χ^2 test, $p \leq 0.05$) and effect size metrics (Cohen's $d \geq 0.5$ for continuous variables; Cramer's $V \geq 0.3$ for categorical variables). This analysis provided an initial data-driven graph linking exposures to outcomes with undirected edges. The literature review stage was then revisited and enriched to decide the direction of each edge between variable with exposure and outcomes and between variables (Figure 8). Finally, domain expertise from a nephrologist was integrated to refine the directed acyclic graphs (DAGs), ensuring biological plausibility and excluding inappropriate variables such as mediators, colliders, and instrumental variables. This integrative approach—combining statistical inference, literature validation, and expert consultation—enabled the construction of reliable causal models for understanding the mechanisms underlying drug-induced AKI.

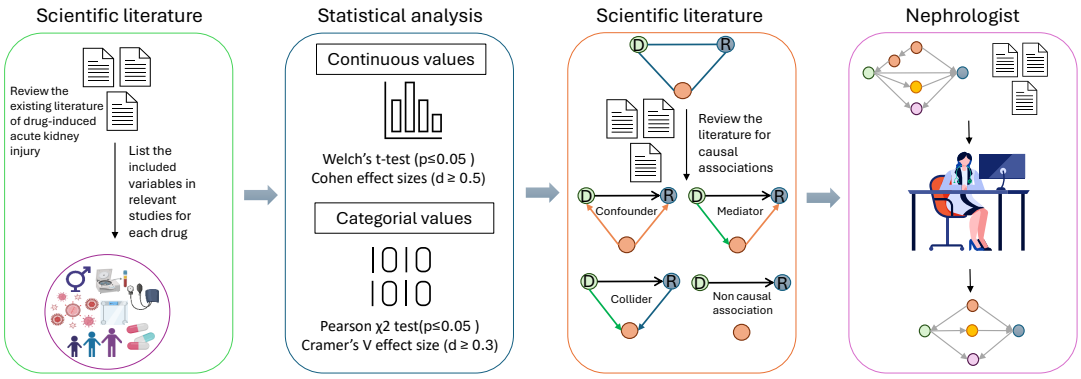

Figure 7: Description of the pipeline for designing DAGs

To determine the causal directionality of relationships within the constructed graphs, a structured synthesis of Google search queries was employed to systematically explore the scientific literature. For each potential edge—whether between exposure (suspect or concomitant drugs) and outcome (acute kidney injury, AKI), or between intermediate variables such as laboratory tests and vital signs—targeted search queries were formulated using combinations of key terms representing both variables and directional phrases. These included expressions such as *"[variable A] affect [variable B]", "[variable A] effect on [variable B]", "[variable A] induced [variable B]", and "[variable A] cause [variable B]"*. The search terms were adapted to capture relationships between drug exposures, physiological measures, and AKI outcomes. Each query result was reviewed to identify evidence supporting or refuting the presumed direction of association. This systematic literature interrogation ensured that every causal edge in the directed acyclic graph (DAG) was

grounded in prior empirical or mechanistic evidence, thereby enhancing the biological and clinical validity of the causal structure. Figure 8 illustrates this procedure in detail.

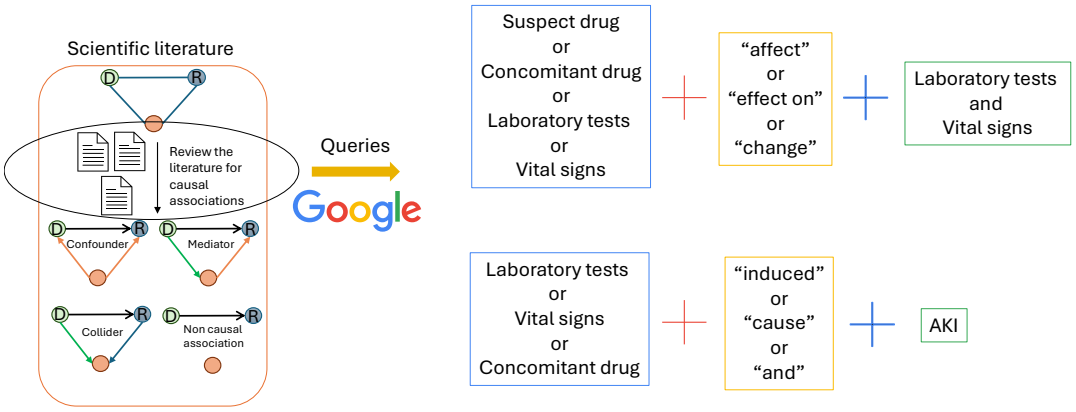

Figure 8: Queries structure for google search to detect the direction of the edges between the variables

Appendix C. Matching

Algorithm 3 performs propensity score matching (PSM) using nearest-neighbor matching with a caliper and a fixed k:1 matching ratio (here $k = 4$). First, the estimated propensity scores are transformed into logit space to stabilize differences between units. A caliper threshold is then computed as a multiple of the standard deviation of the logit scores, restricting matches to treated–control pairs with sufficiently similar propensity scores.

The algorithm identifies treated and control populations, fits a nearest-neighbor search on the control group, and retrieves the k closest controls for each treated unit. A control unit is matched only if it has not been previously used and if its logit distance from the treated unit falls within the caliper. Matched pairs are stored, and the final matched dataset is constructed by combining the selected treated and control units along with their outcomes.

Algorithm 3 Propensity Score Matching with Caliper and k:1 Ratio where $k = 4$

Require: Covariates X, Treatment T, Outcome Y, Propensity scores p, Caliper coefficient c, Ratio k

Ensure: Matched datasets $(X_{\text{match}}, T_{\text{match}}, Y_{\text{match}})$

Compute logit propensity scores:
$\ell_i \leftarrow \log\left(\frac{p_i}{1-p_i+\varepsilon}\right)$
Compute caliper threshold: $\delta \leftarrow c \cdot \text{SD}(\ell)$
Identify treated and control indices: $I_T = \{i : T_i = 1\}, \quad I_C = \{i : T_i = 0\}$
Extract corresponding propensity scores: $p_T \leftarrow p[I_T], \quad p_C \leftarrow p[I_C]$
Fit nearest-neighbor model on controls with k neighbors
Compute neighbor indices: $\text{NN}(i) \leftarrow k$ nearest controls for treated unit i
Initialize matched pair list: $\mathcal{M} \leftarrow \emptyset$
Initialize used controls: $\mathcal{U} \leftarrow \emptyset$
for each treated unit $t \in I_T$ **do**
  for each of the k nearest control candidates $c \in \text{NN}(t)$ **do**
    if $c \in \mathcal{U}$ **then**
      continue to next control
    end if
    if $|\ell_t - \ell_c| \leq \delta$ **then**
      Add (t, c) to $\mathcal{M}$
      Add c to $\mathcal{U}$
    end if
  end for
end for
Extract matched treated and control indices from $\mathcal{M}$: I_T^*, I_C^*
Construct matched datasets:

$$X_{\text{match}} = \begin{bmatrix} X[I_T^*] \\ X[I_C^*] \end{bmatrix}, \quad T_{\text{match}} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}, \quad Y_{\text{match}} = \begin{bmatrix} Y[I_T^*] \\ Y[I_C^*] \end{bmatrix}$$

return $(X_{\text{match}}, T_{\text{match}}, Y_{\text{match}})$

Appendix D. Causal machine learning architectures

S-Learner, shortened from Single-Learner, has been proposed by Hill (2011); Foster et al. (2011). S-Learner uses one model to estimate outcomes for both treated $T_i = 1$ and untreated $T_i = 0$ groups. It works by including the treatment as one of the input variables T_i and then training a single predictive model f. To estimate the treatment effect for an individual i, it calculates the difference between the predicted outcomes $\tau(x_i)$ when the treatment is present ($T_i = 1$) versus when it is not ($T_i = 0$). The CATE is calculated:

$$\tau(x_i) = f(x_i, 1) - f(x_i, 0), \text{ where x is the number of input variables}$$

T-Learner, shortened from Two-Learner, has been discussed in different versions in literature, such as Athey and Imbens (2016); Lu et al. (2018); Powers et al. (2018) who offer some examples

applying decision trees, random forests and gradient boosted trees as base learners, respectively. It builds two separate models, one for the treated group f_1 and one for the untreated group f_0. This setup helps capture any differences between the groups, such as those caused by selection bias or when the relationship between variables behaves differently in each group. The CATE is calculated:

$$\tau(x_i) = f_1(x_i) - f_0(x_i)$$

X-Learner is introduced from Künzel et al. (2019) as an extended form of T-Learner. The calculation of CATE follows three phases, starting like the T-Learner with the estimation of response functions f_1 and f_0, of the treated and untreated groups, respectively (phase 1). In the second phase, the "imported treatment effects", $\tilde{D}_i^1$ and $\tilde{D}_i^0$, are calculated for each group separately by:

$$\begin{aligned} \tilde{D}_i^1 &= Y_i^1 - \hat{Y}_i^{(0)} \quad \text{if } Z_i = 1 \\ \tilde{D}_i^0 &= \hat{Y}_i^{(1)} - Y_i^0 \quad \text{if } Z_i = 0 \end{aligned}$$

Where the group-specific observed outcome, Y_i^Z, and the conditional average potential outcome, $\hat{Y}_i^{(Z)}$, are calculated in phase one. Determine the function for treatment effect, τ_x, by employing two distinct methods: utilize the imported treatment effects as the response variable within the treatment group to derive $\tau_1(x)$ and in the control group for $\tau_0(x)$, respectively. At the last phase (3), the CATE is defined by the weighted average of the two estimates in phase 2, $\pi(x_i)$ is the propensity score (weighting metric). There are also other weighting metrics.

$$\tau(x_i) = \pi(x_i)\tau_0(x_i) + (1 - \pi(x_i))\tau_1(x_i)$$

DML, shortened from Double Machine Learning, was introduced by Chernozhukov et al. (2018). CATE $\theta(x)$, where x is a vector of covariates, is estimated by combining the outcome prediction model and propensity score model into a residual-on-residual regression. DML faces the estimation biases of ML models by applying Neyman orthogonality (Neyman and Scott, 1965) for regularization bias and sample splitting for overfitting bias. The assumed model is partially linear:

$$\begin{aligned} Y &= \theta(X)T + g(X, W) + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X, W] = 0, \\ T &= e(X, W) + \kappa, \quad \mathbb{E}[\kappa \mid X, W] = 0. \end{aligned}$$

where T is the treatment variable, X denotes the feature vector or covariates, W represents additional observed covariates, $g(X, W)$ is an arbitrary function used to predict the outcome variable Y, $e(X, W)$ represents a propensity score model, and $m(X, W)$ denotes a risk prediction model. The terms ε and κ are error coefficients. The dataset is divided into K subsamples, after which $m(X, W) = \mathbb{E}[Y \mid X, W]$ and $e(X, W) = \mathbb{E}[T \mid X, W]$ are calculated within each subsample using arbitrary machine learning models. These nuisance parameters are subsequently utilized to construct a residuals-on-residuals regression model:

$$Y - m(X, W) = \theta(X)(T - e(X, W)) + \varepsilon.$$

The score function ψ is defined as a product of the error term of the residuals-on-residuals regression and the error term of the propensity score model $e(X, W)$:

$$\psi(Z; \theta, h(X, W)) = (Y - m(X, W) - \theta(X)(T - e(X, W))) \cdot (T - e(X, W)).$$

where the observed parameters $Z = Y, T, X, W$ and nuisance parameters $h = m(X, W), e(X, W)$. The estimator $\theta(\hat{X})$ is the solution to:

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \psi \left(Z_i;\ \theta,\ \hat{h}(X_i,\ W_i) \right) = 0$$

The estimated CATE $\hat{\theta}$ minimizes the average of the expected score functions across all K subsamples (Mizuguchi and Sawamura, 2023).

D.1. Evaluation metric for regression used in causal machine learning

This is because these metrics behave differently: MSE and RMSE are minimized in well-performing models (ideally approaching zero), while R^2 is maximized (ideally approaching one). Moreover, MSE and RMSE retain the units of the outcome variable (for example, if the outcome is measured in kilograms, these metrics are also in kilograms), whereas R^2 is proportion of explained variance. To address these differences and create a fair composite score, it is important to ensure that all evaluation metrics have not units of measurement. This is accomplished by applying a logarithmic transformation to the MSE and RMSE values. The final score, which combines all three metrics into a unified performance measure (S), is calculated using the following formula:

$$S = \frac{1}{\log(10 + \text{MSE}) + \log(10 + \text{RMSE}) - 1} + R^2 \tag{1}$$

This formulation allows consistent comparison across models by normalizing the scales of the individual metrics.

D.2. Machine learning algorithms used in causal machine learning and propensity score matching

In each causal machine learning architecture as well as propensity score matching, several ML algorithms are tested as classification or regression models. In S-learner and T-learner only classifications models are used (Logistic Regression, Linear Regression, Stochastic gradient descent, Support Vector Machine, Decision Tree, Random Forest, eXtreme Gradient Boosting, Multilayer Perceptron). For DML architecture only regression models are used (Ridge Regression, Stochastic gradient descent, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, Adaptive Boosting, Extra Trees, Multilayer Perceptron). In X-learner 2 classifiers and 2 regressors are selected based on their performance, as well as one classifier in propensity score matching.

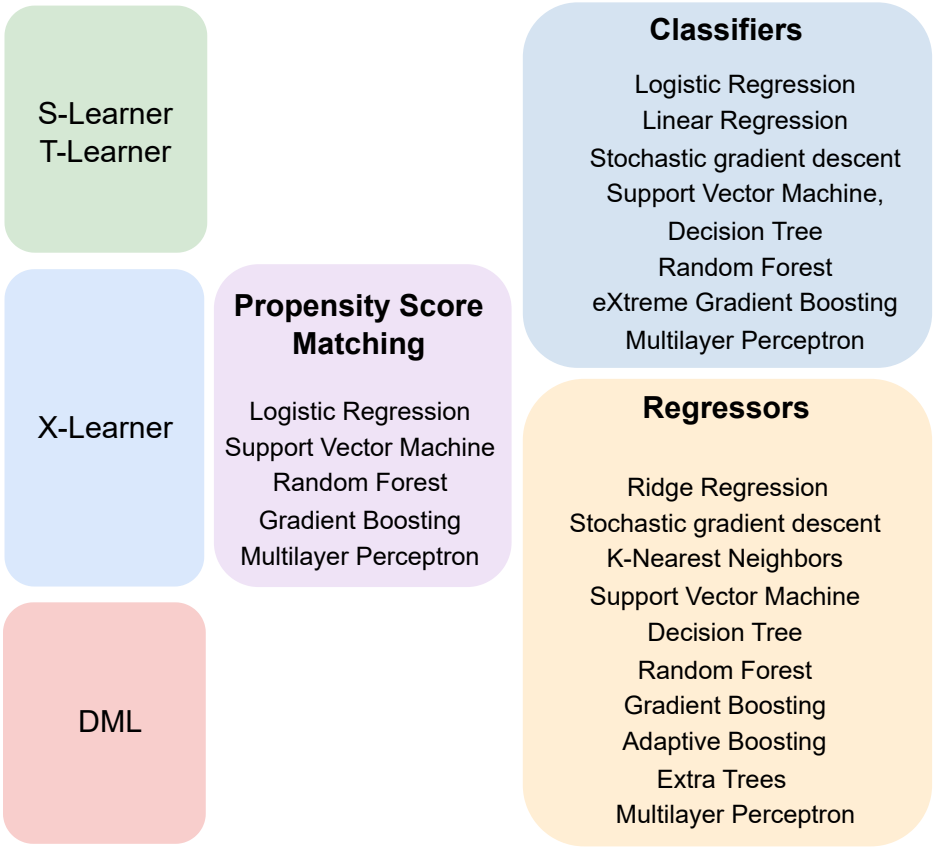

Figure 9: Pipeline for the calculation of causal inference

D.3. Propensity score matching

As X-learner is using propensity score matching (PSM) in its architecture for balance the covariates between the compared treatments (suspect and negative drugs) we tested different ML algorithms (Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, Multilayer Perceptron) and their performance was evaluated with standardize means difference (SMD). It measures covariate balance and it is calculated for each covariate as follows:

$$\text{SMD} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{1}{2}\left(s_T^2 + s_C^2\right)}} \tag{2}$$

where $\begin{cases} \bar{X}_T & = \text{mean of the covariate in the treatment group} \\ \bar{X}_C & = \text{mean of the covariate in the control group} \\ s_T^2 & = \text{variance of the covariate in the treatment group} \\ s_C^2 & = \text{variance of the covariate in the control group} \end{cases}$

Appendix E. Causal assumptions

No Unmeasured Confounding or Ignorability: Suppose that treatment status T is independent of potential outcomes Y in a set of observed covariates X such that

$$Y(t) \perp\!\!\!\perp T \mid X \tag{3}$$

Positivity or Overlap: For all values of X, there is a positive probability of receiving each treatment level:

$$0 < P(T = t \mid X = x) < 1 \tag{4}$$

Consistency: The observed outcome for a unit under treatment $T = t$ is equal to its potential outcome under that treatment:

$$Y = Y(t) \text{ if } T = t. \tag{5}$$

Appendix F. Causal graphs

Table 9 presents the confounders adjusted in each suspect drug base on its DAGs (nephrologist, literature) architecture. In every suspect drug graph the existance of AKI stage (> 0) before the drug administration is a common confounder and is mentioned as *Pre AKI*.

Drug (DAG agreement)	Literature	Nephrologists
Ibuprofen (No)	Age, Anti-inflammatory and Antirheumatic products, Antihypertensives, Charlson comorbidities, Contrast media, Corticosteroids, Diabetes drugs, Diuretics, Gender, Topical products for joint and Muscular pain, Weight, Pre AKI	Age, Anti-inflammatory and Antirheumatic products, Antihypertensives, Charlson comorbidities, Contrast media, Corticosteroids, Diuretics, Gender, Topical products for joint and Muscular pain, Weight, Pre AKI
Ketorolac (No)	Age, Antibacterials, Antihypertensives, Charlson comorbidities, Contrast media, Corticosteroids, Diuretics, Gender, Weight, Glucose, Pre AKI	Age, Antibacterials, Antihypertensives, Charlson comorbidities, Contrast media, Corticosteroids, Diuretics, Gender, Weight, Sodium, Pre AKI
Vancomycin (Yes)	Age, Antibacterials, Antihypertensives, Charlson comorbidities, Contrast media, Corticosteroids, Diuretics, Gender, Vaccines, Weight, Pre AKI	Age, Antibacterials, Antihypertensives, Charlson comorbidities, Contrast media, Corticosteroids, Diuretics, Gender, Vaccines, Weight, Pre AKI
Lisinopril (No)	Age, Antibacterials, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI	Age, Agents Acting on the Renin-Angiotensin System, Antibacterials, Antiepileptics, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI
Furosemide (Yes)	Age, Antibacterials, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI	Age, Antibacterials, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI
Pantoprazole/ Omeprazole (Yes)	Age, Antibacterials, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI	Age, Antibacterials, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI
Allopurinol (No)	Age, Agents Acting on the Renin-Angiotensin System, Antibacterials, Charlson comorbidities, Contrast media, Diuretics, Gender, Weight, Pre AKI	Age, Agents Acting on the Renin-Angiotensin System, Antibacterials, BUN, Bicarbonate, Blood Pressure, Charlson comorbidities, Chloride, Contrast media, Creatinine, Drugs for Acid-related disorders, Gender, Immunosuppressants, Potassium, Psycholeptics, Sodium, Pre AKI

Table 9: Confounders for each drug related to AKI according to the DAGs in literature and nephrologist review

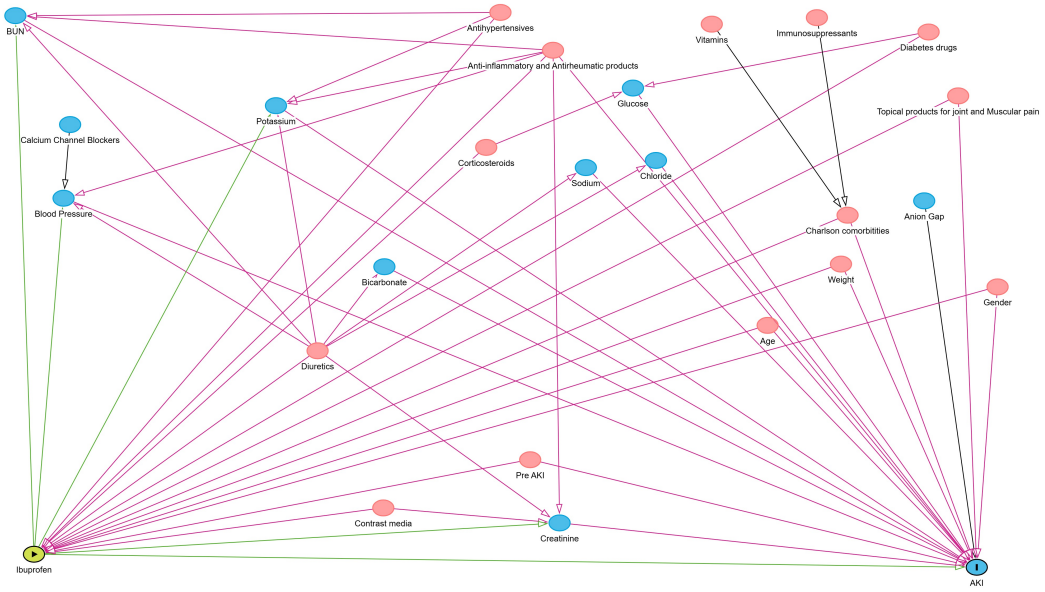

Figure 10: DAG architecture designed based on the literature review for Ibuprofen

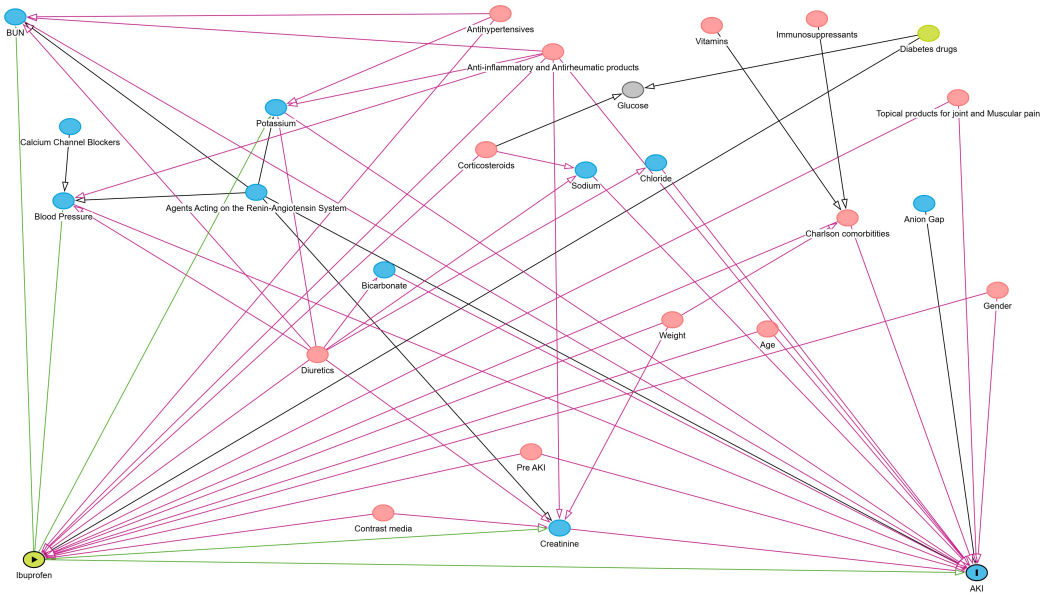

Figure 11: DAG architecture designed based on the nephrologist review for Ibuprofen

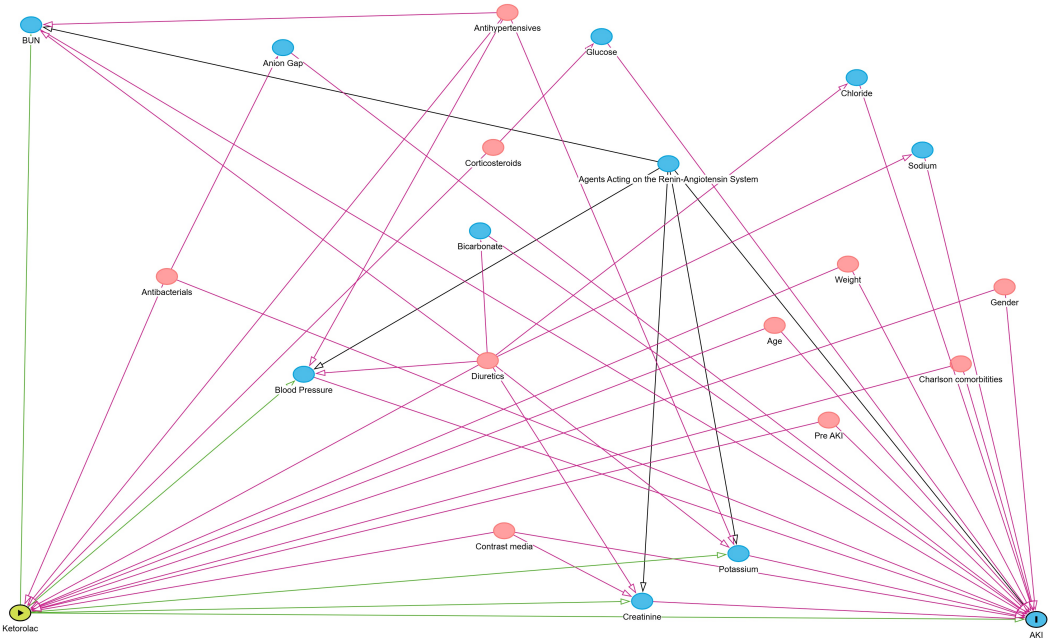

Figure 12: DAG architecture designed based on the literature review for Ketorolac

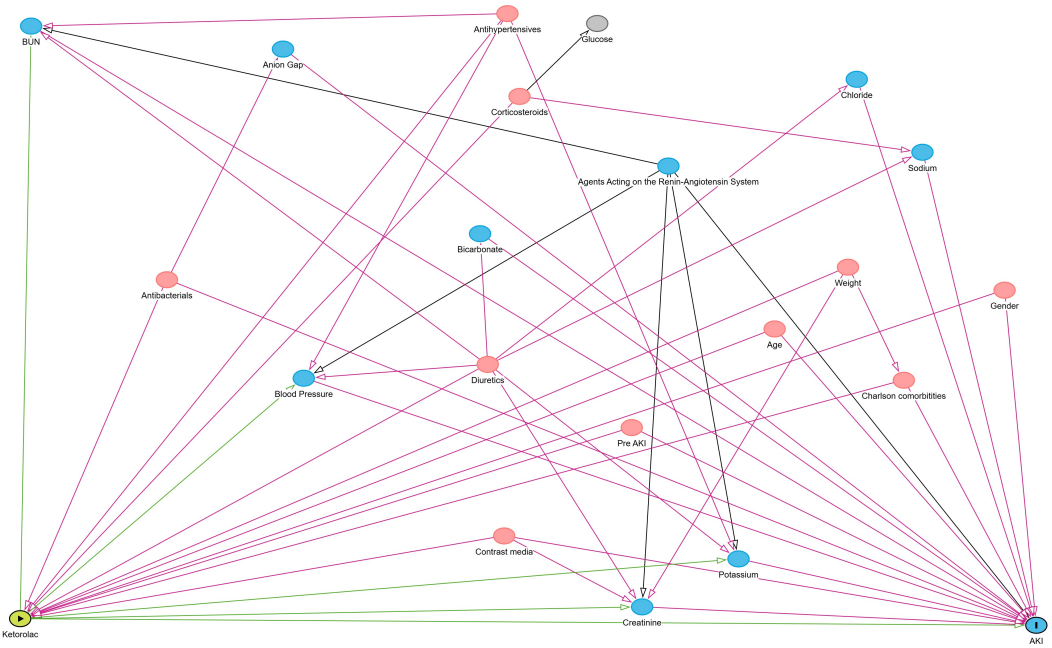

Figure 13: DAG architecture designed based on the nephrologist review for Ketorolac

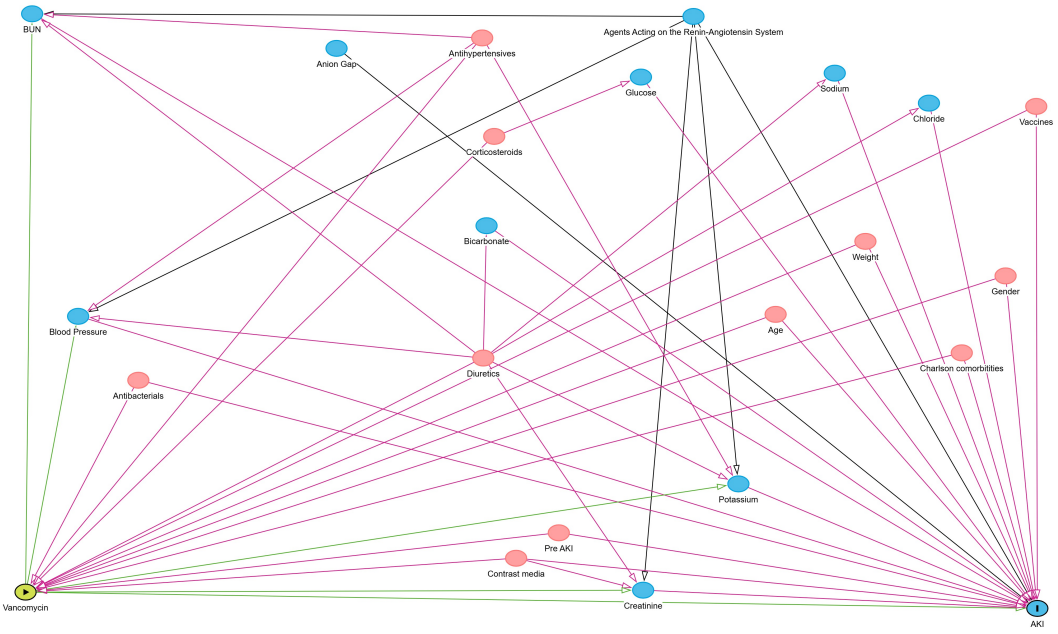

Figure 14: DAG architecture designed based on the literature and nephrologist review for Vancomycin

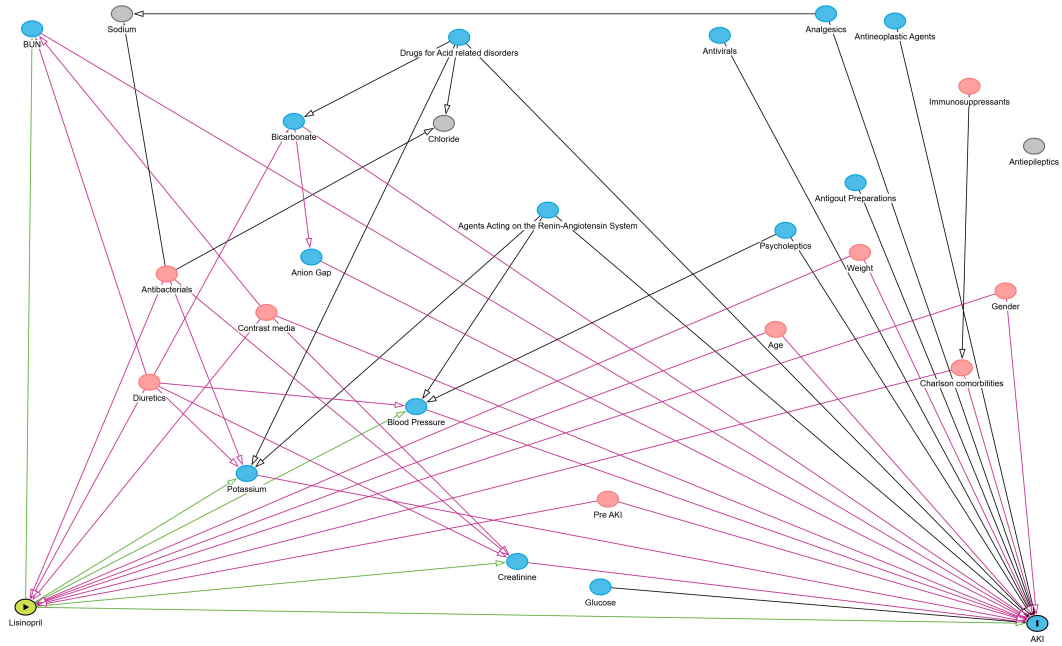

Figure 15: DAG architecture designed based on the literature review for Lisinopril

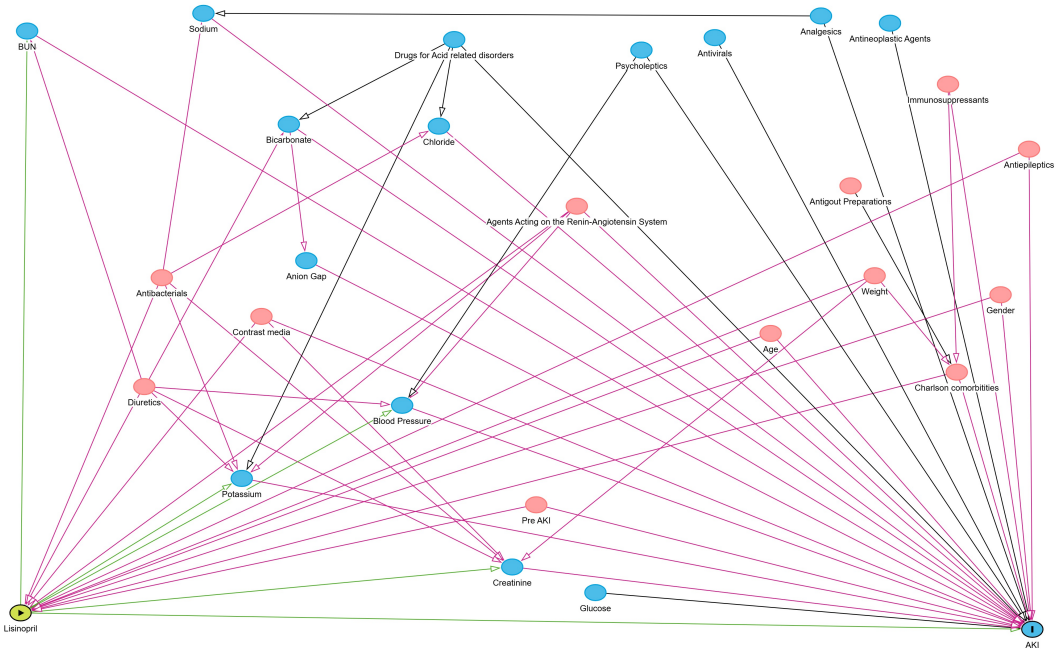

Figure 16: DAG architecture designed based on the nephrologist review for Lisinopril

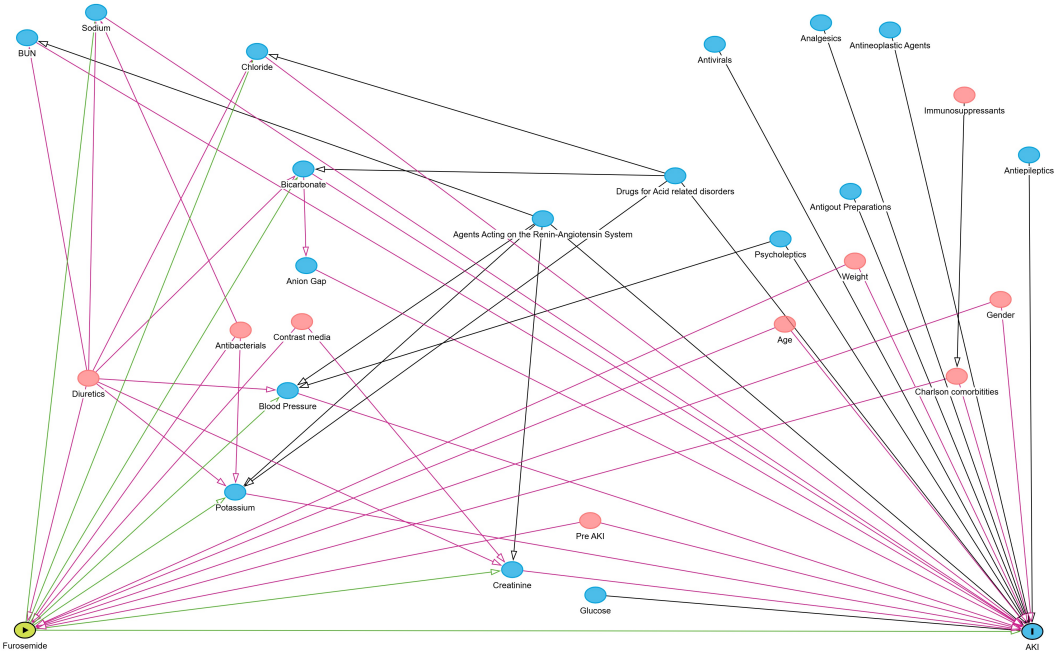

Figure 17: DAG architecture designed based on the literature and nephrologist review for Furosemide

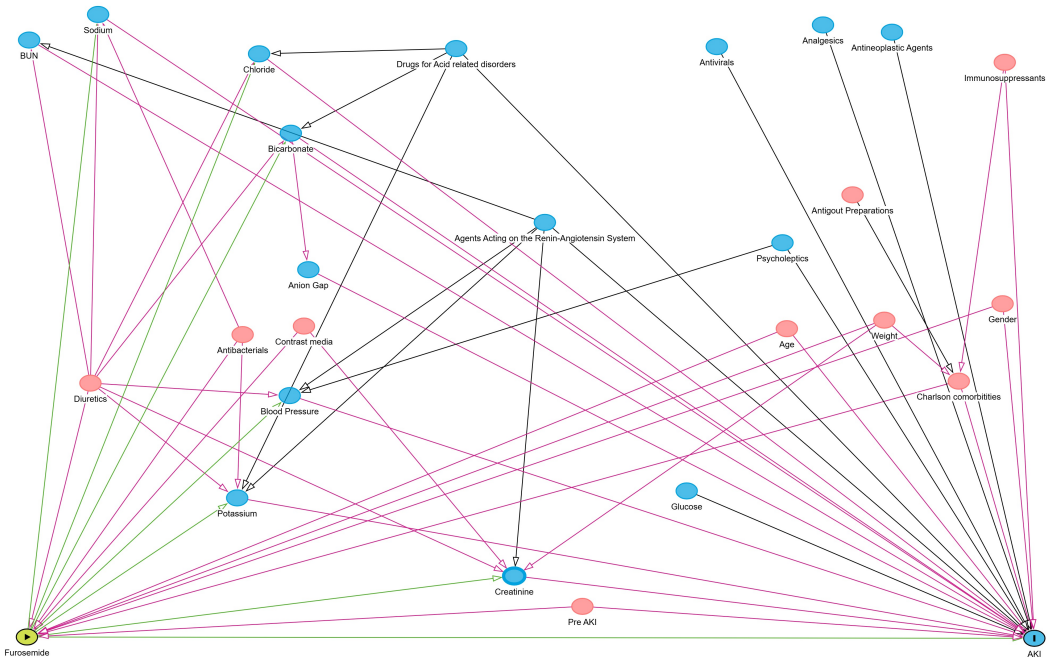

Figure 18: DAG architecture designed based on literature and nephrologist for Furosemide

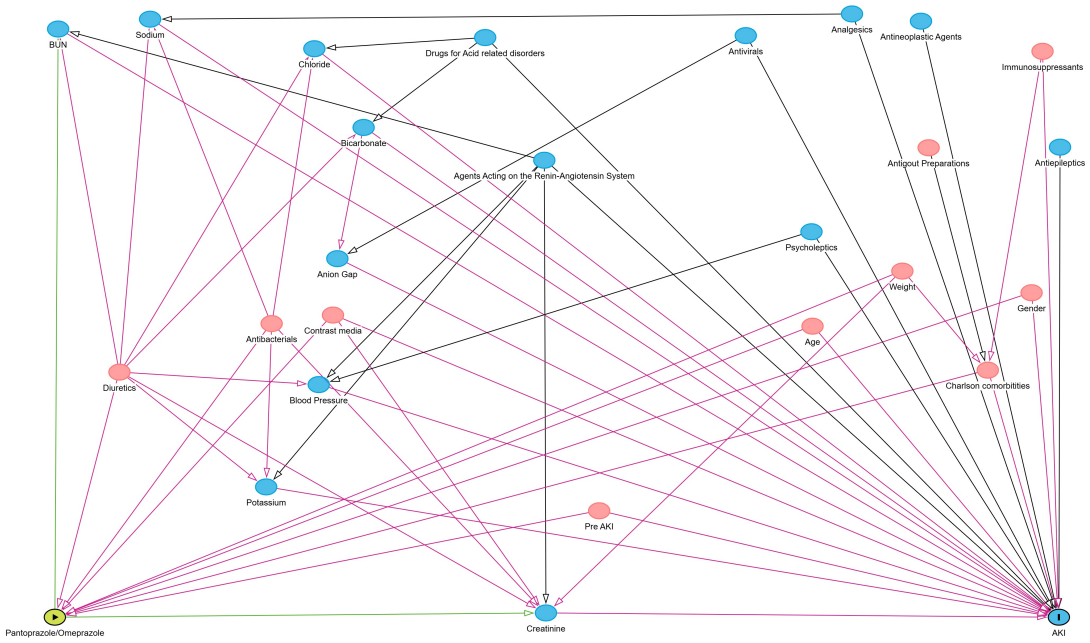

Figure 19: DAG architecture designed based on literature and nephrologist for Pantoprazole/Omeprazole

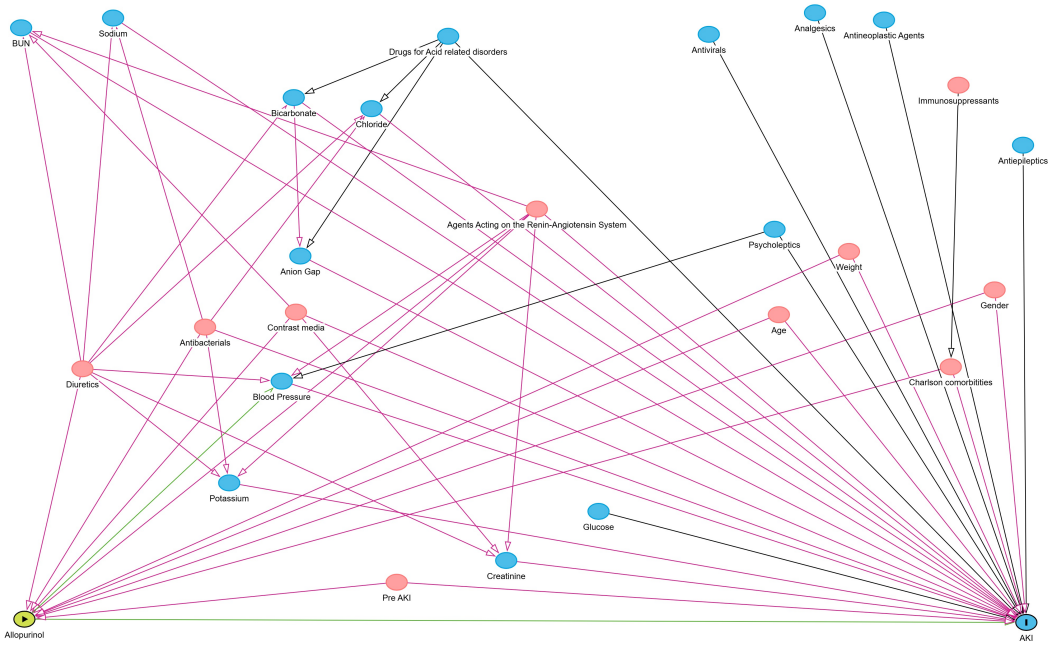

Figure 20: DAG architecture for allopurinol-induced AKI designed based on the literature review

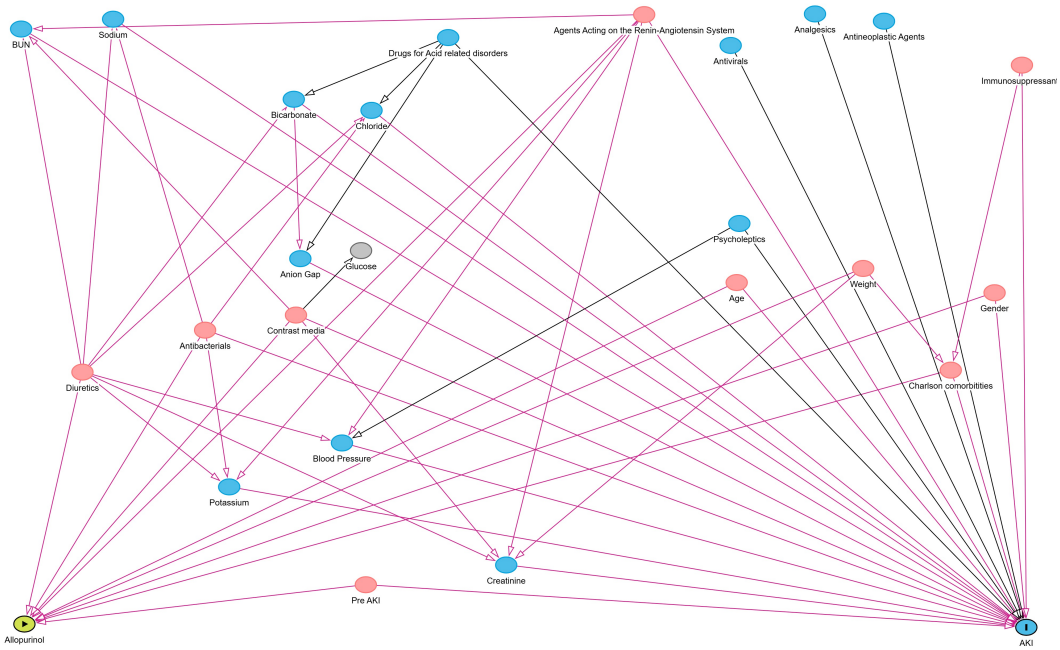

Figure 21: DAG architecture for allopurinol-induced AKI designed based on the neprhologist knowledge

Appendix G. Matching

Table 10: Best machine learning models for propensity score matching separated into the two confounder sets of the literature review and nephrologist, where laboratory tests are represented as mean values - Link to github.

Table 11: Best machine learning models for propensity score matching separated into the two confounder sets of the literature review and nephrologist, where laboratory tests are represented as first and last value - Link to github.

Appendix H. Best machine learning models

The Random Forest classifier consistently outperformed all other ML models in the S-learner architecture across all suspect drugs and control drug combinations. It achieved good performance in terms of accuracy ($\geq$ 0.99), precision, recall, and F1 in ibuprofen, ketorolac, vancomycin, furosemide, pantoprazole, omeprazole, and allopurinol where F1 score range form 89% to 96%. Notably, Random Forest reached the lowest F1 score of 77% for lisinopril when paired with Lactulose. The ML models performance was calculated with cross validation method. Table 12 presents a summation of the results.

Drug	**Best F1**	**Model**	**Negative drug**
Ibuprofen	0.96	Random Forest	Lactulose
Ketorolac	0.94	Random Forest	Lactulose
Vancomycin	0.96	Random Forest	Lactulose / Prochlorperazine
Lisinopril	0.77	Random Forest	Lactulose
Furosemide	0.89	Random Forest	Simethicone
Pantoprazole	0.92	Random Forest	Simethicone
Omeprazole	0.93	XGBoost	Lactulose / Prochlorperazine
Allopurinol	0.9	Random Forest	Lactulose

Table 12: Best-performing models for each drug using S-learner methodology

Performance of models in the T-learner framework was more heterogeneous. While Random Forest and XGBoost still emerged frequently as high performers, a broader set of classifiers including MLP Classifier, Decision Tree, and Stochastic Gradient Descent (SGD) Classifier were also used effectively. More specifically, ibuprofen had optimal performance using dual Random Forests (F1 = 0.95) or XGBoost (F1 = 0.96) and poorer performance with SDG classifier (F1=0.82) depending on the control group. Ketorolac succeeded the best perfrormance with Random forest (F1=0.94) and slightly lower performance with XGBoost (F1=0.92) and MLP (F1=0.92). Vancomycin achieved its best performance in combining SGD Classifier with XGBoost (F1=0.92) or

Random Forest (0.92). Despite generally lower metrics compared to S-learners, T-learners demonstrated moderate to high performance (F1 scores between 0.85–0.96) and may offer advantages in modeling treatment heterogeneity where group-specific outcome modeling is beneficial, as this separation allows each model to learn the patterns and covariate interactions that are specific to its group. In lisinopril, the combination of XGBoost and Random Forest achieved the best performance with F1 score equal 0.83 (simethicone, prochlorperazine). In furosemide, Decision Tree and MLP Classifier in simethicone as control drug achieved the highest F1 score equals 0.87. Pantoprazole achieved the highest F1 score with Decision tree and SDG classifier combination (0.87). The ML models with the best F1 score in Pantoprazole achieved with the combination of Decision Tree and SGD Classifier with prochlorperazine control drug. Decision tree was the best ML algorithm in the first model of omeprazole and its combination with MLP in lactulose control drug performed slightly better in recall (0.87) than the others (0.86). Accuracy, precision and F1- score are the same in other control drugs, 88, 86, 86, respectively. In allopurinol, the combination of two XGBoost models in simethicone control drug have the best results (F1=0.88). Table 13 presents the best models.

Drug	**Best F1**	**Model Combination**	**Negative drug**
Ibuprofen	0.96	XGBoost, XGBoost	Lactulose
Ketorolac	0.94	Random Forest, Random Forest	Lactulose
Vancomycin	0.92	Random Forest, SGD Classifier	Lactulose
Lisinopril	0.83	XGBoost, Random Forest	Simethicone / Prochlorperazine
Furosemide	0.87	Decision Tree, MLP Classifier	Simethicone
Pantoprazole	0.87	Decision Tree, SGD Classifier	Simethicone / Prochlorperazine
Omeprazole	0.86	Decision Tree, Random Forest	Simethicone / Prochlorperazine
Allopurinol	0.88	XGBoost, XGBoost	Simethicone

Table 13: Best-performing model combinations per drug using T-learner methodology

X-learners, combining classification and regression stages, exhibited moderate model complexity and competitive performance across suspect drugs. Furthermore a PSM model is required which is selected based on the matching results presented above. Models involved as classifiers were Support Vector Classifiers, Random Forest, Decision Trees, SGD Classifier, MLP Classifier and XGBoost. In terms of regressors, MLP Regressor, Support Vector Regressor, Random Forest, Decision Tree and Gradient Boosting developed the most accurate models. The datasets with simethicone as control drug achieved the best overall score of ML models. In particular, ibuprofen had the best performance with a combination of MLP and SVM classifiers and 2 MLP regressors in the dataset with lactoluse control drug (overall score = 2.17). Same as ibuprofen, ketorolac dataset with lactoluse succeed the best results (overall score=2.33) with the combination of Support Vector and SGD Classifiers and 2 Support Vector Regressors. Vancomycin with simethicone had score 2.63 in XGBoost and SGD Classifiers with 2 Random forest regressors. Furosemide (Random Forest and Support Vector Classifiers with 2 Random forest regressors) and patnoprazole (Decision Tree and Support Vector Classifiers with 2 Decision Tree regressors) had the highest overall scores with simethicone control drug, 2.64 and 2.68, respectively. Lisinopril had the same overall score 2.63 and the same combination of models for each control drug (Random Forest and Support Vector Classifiers, 2 Random Forest regressors). Omeprazole and allopurinol ad the highest overall scores with prochlorperazine control drug, 2.65 and 2.20, respectively. Table 14 summarizes the results.

Drug	Performance	Model Combination	Negative drug
Ibuprofen	2.17	SGD Classifier, SGD Classifier	Lactulose
Ketorolac	2.33	Support Vector Classifier, SGD Classifier, Support Vector Regressor, Support Vector Regressor	Lactulose
Vancomycin	2.63	XGBoost, SGD Classifier, Random Forest, Random Forest	Simethicone
Lisinopril	2.42	Random Forest, Support Vector Classifier, Random Forest, Random Forest	Any (All Equal)
Furosemide	2.64	Random Forest, Support Vector Classifier, Random Forest, Random Forest	Simethicone
Pantoprazole	2.68	Decision Tree, Support Vector Classifier, Decision Tree, Decision Tree	Simethicone
Omeprazole	2.65	Decision Tree, Support Vector Classifier, XGBoost, XGBoost	Prochlorperazine
Allopurinol	2.20	Random Forest, Support Vector Classifier, Decision Tree, Decision Tree	Prochlorperazine

Table 14: Best-performing model combinations per drug using X-learner methodology

DML architecture inludes 3 regressors, one for the outcome model, one for the treatment model, and one for final estimator. Gradient Boosting and Decision Trees dominated across the modeling stages (outcome, treatment, and final estimator). The datasets with lactulose as control drug achieved the best overall score of ML models (Table 15). More precisely, in ibuprofen (XGBoost, Random Forest, Decision Tree), ketorolac (Gradient Boosting, XGBoost, Decision Tree) and lisinopril (Gradient Boosting, Gradient Boosting, Decision Tree) models performed better in lactulose than the other control drugs with 1.95, 1.92 and 1.63, respectively. Furthermore, pantoprazole (Gradient Boosting, Decision Tree, Decision Tree) and furosemide (Gradient Boosting, Gradient Boosting, Decision Tree) succeeded the best overall score performance with simethicone as the control drug, 1.76 and 1.93, respectively. Omeprazole da the mest performance with prochlorperazine as control drug and overall S score 1.92. Vancomycin achieved the highest S overall score equal to 1.96 in both prochlorperazine and lactulose datasets. Allopurinol had the highes S overall score of 1.76 in simethicone and lactulose.

Appendix I. Conditional Average Treatment Effect with matching

In the datasets after matching the number of patients significantly decreased. Thus, the calculation of CATE produce wide confidence intervals (CI width $\simeq$ 1). Figure 22 illustrates the methodology pipelines are mostly used for the production of valid CATEs (CI width leq 0.5, we select this width because it is the largest CI width among significant CATEs whose intervals exclude zero.) from matched samples. Vancomycin produced the most valid results. Prochlorperazine as negative control drug with nephrologist DAGs, representation of laboratory values as mean and X-learner succed the most valid CATEs in 7 out of 8 suspect drugs. In contrast, allopurinol's CIs width $\simeq$ 1 so the CATE results are not reliable. Ketorolac, lisinopril, vancomycin and furosemide with X-learner method produced statistical significant results.

Drug	Model Combination	Performance	Negative drug
Ibuprofen	XGBoost, Random Forest, Decision Tree	1.95	Lactulose
Ketorolac	Gradient Boosting, XGBoost, Decision Tree	1.92	Lactulose
Vancomycin	Gradient Boosting, Gradient Boosting, Decision Tree	1.96	Prochlorperazine / Lactulose
Lisinopril	Gradient Boosting, Gradient Boosting, Decision Tree	1.63	Lactulose
Furosemide	Gradient Boosting, Gradient Boosting, Decision Tree	1.76	Simethicone
Pantoprazole	Gradient Boosting, Decision Tree, Decision Tree	1.93	Simethicone
Omeprazole	Gradient Boosting, XGBoost, Decision Tree	1.92	Prochlorperazine
Allopurinol	Gradient Boosting, Gradient Boosting, Decision Tree	1.76	Simethicone / Lactulose

Table 15: Best-performing model combinations per drug using DML-learner methodology

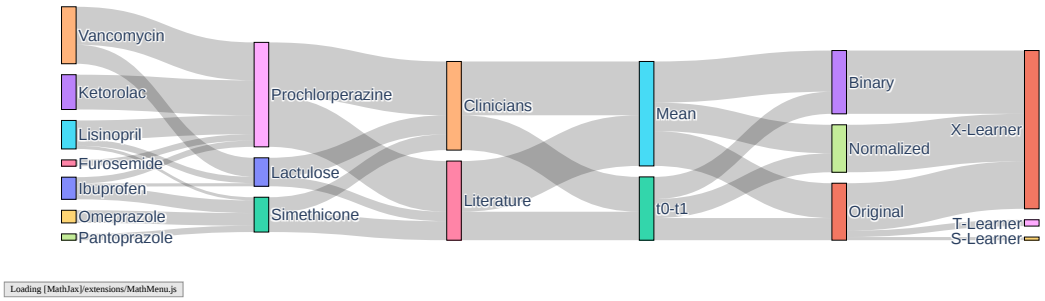

Figure 22: CATE Estimates with confidence inderval (CI) width ≤ 0.5 of matched patients

For ibuprofen (CATE mean: -0.04, SD: 0.03), ketorolac (CATE mean: 0.33, SD: 0.09), lisinopril (CATE mean: 0.61, SD: 0.1), furosemide (CATE mean: 0.49, 1 valid resutl) and vancomycin (CATE mean: 0.68, SD: 0.12) predict the narrowest CIs with binary representation of the concomitant drugs and pantoprazole (CATE mean: 0.07, 1 valid resutl) and omeprazole (CATE mean: 0.02, 1 valid resutl) with original datasets.

Appendix J. Conditional Average Treatment Effect without matching

J.0.1. IBUPROFEN

After filtering the results, in the remained CML methods, where the laboratory tests represent as mean values (Figure 23), S-learner has the most valid results (15/18), DML and X-learner have the

same valid CATEs (12/18) and T-learner has the least accepted CATEs (10/18). Effect estimates from multiple meta-learners and different review perspectives were generally close to zero with narrow confidence intervals, indicating no strong evidence for a causal effect of ibuprofen on AKI compared to the selected controls under the presented analysis conditions. Specifically, when using simethicone as the control, the CATE estimates derived from literature-based meta-learners were very close to zero (CATE range 0 to 0.0021) with confidence intervals overlapping zero. Similarly, nephrologist-reviewed estimates were also near zero (CATE range 0 to 0.0003), reinforcing the absence of a detectable causal effect in this comparison.

When lactulose was the control, only one valid result exists based on the DAG created from the literature review. The few available estimates were uniformly zero with no significant deviations. For prochlorperazine, some estimates indicated slight negative effects with wide confidence intervals crossing zero (e.g., nephrologist DAG, DML-learner, CATE = -0.0152, CI: -0.064 to 0.034), suggesting uncertainty in the direction and magnitude of any effect. Notably, the normalized input data with prochlorperazine control showed moderate positive CATE values around 0.31–0.33 with CI excluding zero (e.g., literature DML-learner CATE = 0.329 (0.236–0.422); nephrologist DML-learner CATE = 0.310 (0.210–0.409)), which may indicate a potential increased risk signal in this analytic scenario. However, other learner methods and DAG perspectives yielded CATEs near zero or with CI including zero, emphasizing variability dependent on model and DAGs.

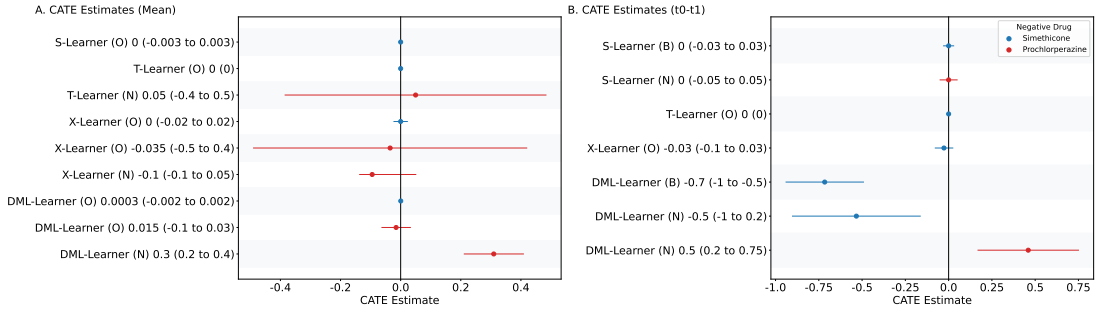

Figure 23: CATE Estimates (with 95% CI) for Ibuprofen

In the valid CML methods where the results are represented as first and last value (Figure 23), S-learner has the most valid results (12/18), X-learner has 10 CATEs, X-learner has 8 CATEs, and T-learner has the least accepted CATEs (6/18). In the ibuprofen drug, there are not valid results against lactulose. A consistent signal was observed across several learners, suggesting a potential protective association between Ibuprofen and AKI, especially in DML-based analyses. Both nephrologist- and literature-based inputs yielded negative CATEs, indicating that Ibuprofen was associated with lower rates of AKI compared to Simethicone, though the effect size varied. Furthermore, several analyses using S-, T-, and X-learners across the datasets consistently yielded null effects (CATE = 0) with narrow CIs centered around zero, indicating that there was no detectable difference in AKI incidence between Ibuprofen and Simethicone.

Contrary to the results with Simethicone, the CATEs for Ibuprofen compared with Prochlorperazine leaned positive, indicating a potential increased risk of AKI associated with Ibuprofen in this comparison. This result was particularly pronounced in the DML learners using normalized inputs. The findings were consistent across literature- and nephrologist-informed models. This

suggests a moderate to strong association of higher AKI rates with Ibuprofen when compared to Prochlorperazine. In contrast to DML, S-learner predicted a zero CATE with narrow CIs.

The DML learners appeared to be more sensitive and discriminative, detecting both negative CATEs (Ibuprofen vs. Simethicone) and positive CATEs (Ibuprofen vs. Prochlorperazine) with narrow and informative confidence intervals (positive values have slightly narrower CI than negative). The null results reinforce the robustness of the effect patterns only observed with DML-based learners. The divergence in directionality of the CATE across different control drugs underscores the importance of control selection and potential underlying confounding. The data do not suggest a uniform risk profile for Ibuprofen across comparators.

J.0.2. KETOROLAC

In the valid CML methods in mean values of laboratory tests (Figure 24), X-learner results are all valid (18/18), S-learner has fewer than a half valid CATEs than X-learner (7/18), T-learner has a small number of accepted CATEs (3/18) and DML has only one valid result. When lactulose serves as the control, results show a wider spread of CATEs, ranging from modest positive to negative values. The literature-based X-learner on normalized inputs estimated a CATE of 0.19 (95% CI: -0.02 to 0.40), suggesting a possible increased risk, but again with confidence intervals including zero, which defines CI as insignificant. Nephrologist estimates tended to be lower and more conservative, e.g., 0.07 (-0.20 to 0.35) for the normalized X-learner. Some analyses even show slight protective associations (negative CATEs), such as the literature's binary X-learner estimate of -0.07 (-0.44 to 0.30). The wide confidence intervals reflect substantial uncertainty likely related to sample sizes and inherent variability. Overall, no definitive conclusion on risk can be drawn between ketorolac and lactulose.

The CATE estimates comparing ketorolac to simethicone suggest a slight positive association with acute kidney injury (AKI), but with confidence intervals spanning zero and extending into both negative and positive ranges, indicating statistical uncertainty. Both literature and nephrologist assessments, across various meta-learner models (X-learner, S-learner) and data transformations (original, binary, normalized), yield comparable results. The largest point estimates were observed using the X-learner on the original input, with nephrologists estimating an CATE of 0.16 (95% CI: -0.10 to 0.42) and literature reporting 0.14 (-0.12 to 0.39). These findings do not conclusively demonstrate an increased risk of AKI attributable to ketorolac relative to simethicone.

Comparisons with prochlorperazine consistently show CATE estimates near zero with narrow confidence intervals crossing zero, suggesting no significant difference in AKI risk between ketorolac and prochlorperazine. Both nephrologist and literature evaluations agree closely, with estimates such as 0.05 (-0.11 to 0.21) for literature and -0.02 (-0.17 to 0.14) for nephrologists using the original input with X-learner. Binary and normalized inputs yield even smaller effects. These findings imply that ketorolac's AKI risk is comparable to that of prochlorperazine.

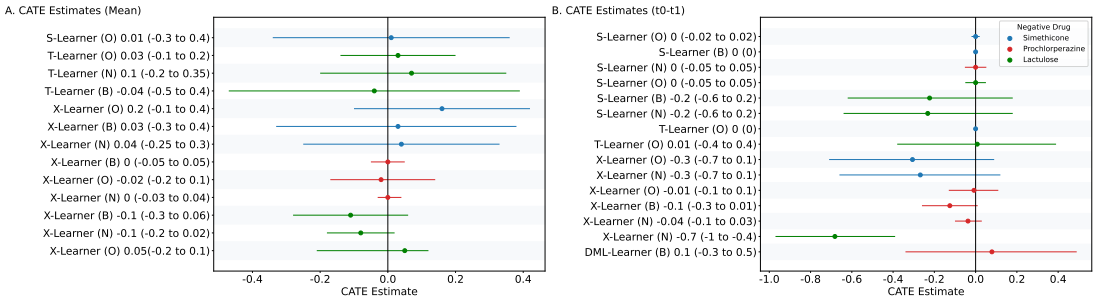

Figure 24: CATE Estimates (with 95% CI) for Ketorolac

In ketorolac's datasets with laboratory tests as first and last value (Figure 24), X-learner includes the most CATEs (13/18), S-learner with 12 CATEs, T-learner with only 3 accepted CATEs, and DML with one. The estimated CATEs for Ketorolac vs. Simethicone largely suggest either null effects or a potentially protective association (i.e., lower AKI risk with Ketorolac), particularly in literature-based reviews. Multiple T- and S-Learner models reported zero CATEs with narrow CIs, indicating high stability but no detectable effect (e.g., T-Learner Original: CATE equals 0 with CI upper and lower equals to 0). Literature-based X-Learner using original inputs yielded a CATE of -0.365 with a lower bound reaching -0.72 and an upper bound close to zero (-0.01), suggesting a modestly protective effect.

The CATEs from the Ketorolac comparison with Prochlorperazine mostly fall close to zero with wide and overlapping confidence intervals, indicating no statistically significant association. Both literature and nephrologist-based reviews using X-Learner and DML models produce CATEs ranging from -0.124 to 0.079, all with CIs that include 0. These results suggest no conclusive evidence of increased or decreased AKI risk from Ketorolac when compared to Prochlorperazine.

Results suggest a significant increase in AKI risk associated with Ketorolac compared to Lactulose across both nephrologist- and literature-based reviews. These effects are consistently negative with tight confidence intervals. The strongest signal is found using the X-Learner with normalized inputs, with CATEs of -0.609 (CI: -0.94, -0.28) for literature-based and -0.682 (CI: -0.97, -0.39) for nephrologist-based inputs. These results suggest robust evidence of not ketorolac-induced AKI.

J.0.3. VANCOMYCIN

Vancomycin has 7 valid CATE results (laboratory tests as mean values) out of 18 with S-learner method. Three comparisons between Vancomycin and Lactulose were conducted, using original, binary, and normalized input formats, respectively. In all scenarios, the S-Learner model was applied. The estimated CATEs consistently indicated a slight positive association between Vancomycin exposure and AKI risk; however, all confidence intervals were wide and included zero, suggesting that the observed effects were not statistically significant. The wide confidence intervals and overlap with the zero suggest the observed differences could be attributed to random variation rather than a true treatment effect (Figure 25).

There were no results presented for Vancomycin in comparison with Simethicone or Prochlorperazine that met the inclusion criteria for reporting (i.e., confidence intervals did not fall within the predefined thresholds).

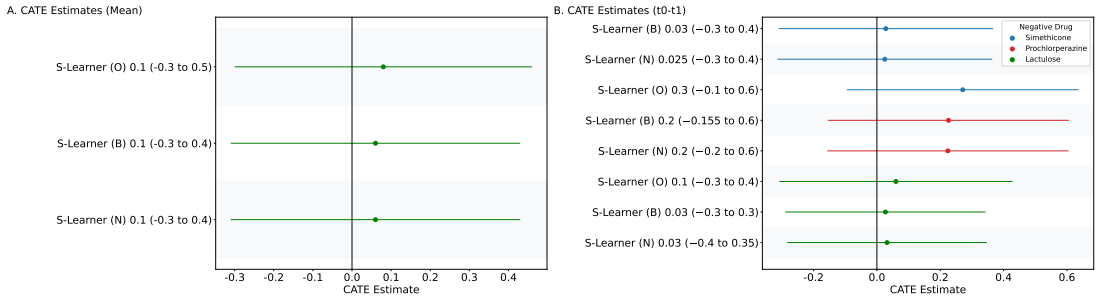

Figure 25: CATE Estimates (with 95% CI) for Vancomycin

Only S-learner, in the vancomycin datasets with laboratory tests as first and last value, produced 9 valid CATE results out of 18. When Vancomycin was compared to Simethicone, the estimated CATEs were consistently low, ranging from 0.025 to 0.028 across preprocessing types. All CIs spanned zero, suggesting no statistically significant causal association between Vancomycin use and increased risk of AKI in comparison to Simethicone. These findings imply that the background rate of AKI is likely comparable between the Vancomycin and Simethicone-exposed populations when adjusted for observed confounders.

Prochlorperazine produced moderately high CATE estimates when compared to Vancomycin. The CATEs ranged from 0.2239 to 0.2711, indicating a potentially increased risk of AKI associated with Vancomycin use. However, all CIs still included zero (e.g., -0.095 to 0.637), and thus, the association remains statistically inconclusive. The consistency of positive CATEs across preprocessing methods, however, may suggest a trend worth investigating further in larger or more targeted studies.

In the comparison with Lactulose, CATE values were similarly small, ranging from 0.0269 to 0.0598 across different feature processing strategies. Again, all CIs encompassed zero. While the direction of the CATEs suggested a slightly increased risk of AKI with Vancomycin, the magnitude was minimal and not statistically significant. This suggests that although there may be a slight trend, the data do not provide strong evidence to support a meaningful causal relationship in this comparison.

J.0.4. LISINOPRIL

In the case of lisinopril (laboratory tests as mean values), unlike X-learner and S-learner, which do not produce any invalid results, DML yields only one valid result (Figure 26). Across all input types, results consistently suggested a negative CATE, indicating a lower risk of AKI for patients exposed to Lisinopril compared to Simethicone. The most substantial negative effects were estimated using X-Learner across all input types. For instance, with binary input and nephrologist-based confounding adjustment, the CATE was -0.23 (95% CI: -0.46, -0.00), which approached statistical significance.

Similar to Simethicone, comparisons with Prochlorperazine showed predominantly negative CATEs, again suggesting a possible protective effect of Lisinopril. X-Learner models yielded the strongest negative effects, particularly for normalized input with literature-based confounding (CATE: -0.24, 95% CI: -0.47, -0.01).

In contrast to the other control drugs, CATE estimates from comparisons with Lactulose were closer to zero and less consistent. Most S-Learner models indicated small, statistically insignificant positive effects, while X-Learner models leaned toward slight negative effects. The closest to a potentially meaningful result was observed with normalized inputs and nephrologist-based X-Learner estimates (CATE: -0.17, 95% CI: -0.42, 0.08), though still not statistically significant.

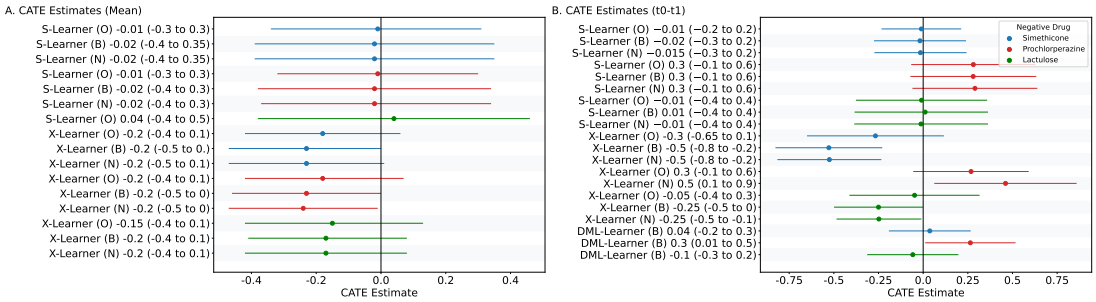

Figure 26: CATE Estimates (with 95% CI) for Lisinopril

In lisinopril's datasets where laboratory tests represent as first and last value (Figure 26) S-learner has only valid CATEs (18/18). X-learner has 16 CATEs and T-learner and DML have 3 and 1 valid results, respectively. The estimated CATEs with Simethicone as a control show negative or near-zero values, particularly significant when using the X-Learner and Normalized/Binary inputs, indicating a reduced AKI risk with Lisinopril, though some CI intervals suggest high uncertainty.

The estimated CATEs comparing Lisinopril to Prochlorperazine consistently indicate a positive effect, suggesting a higher risk of AKI associated with Lisinopril. This trend holds across all input formats and meta-learners, with most confidence intervals not crossing zero or having upper bounds well below 1.

CATEs with Lactulose as the control are generally small and negative, with some statistically significant findings in the X-Learner configurations. These results suggest a minimal to modest reduction in AKI risk with Lisinopril use.

Prochlorperazine consistently shows positive CATEs across models and data representations, suggesting a higher risk of AKI associated with Lisinopril exposure. In contrast, Simethicone exhibits negative or near-zero CATEs, with statistically significant negative values observed under the X-Learner framework. This pattern may point to a potential protective effect of Simethicone or highlight confounding factors influencing the result. Meanwhile, Lactulose demonstrates small, consistently negative CATEs, with statistically significant results under the X-Learner. These findings suggest a slight reduction in AKI risk with Lisinopril relative to Lactulose; however, the magnitude of the effect is modest and likely of limited clinical relevance.

J.0.5. FUROSEMIDE

Only X-learner (2/18) and S-learner (3/18) have valid CATE values in the mean laboratory values' dataset in furosemide (Figure 27). In the comparison between Furosemide and Simethicone, only one model met the reporting criteria. Using binary input and the S-Learner, the estimated CATE was 0.01 (95% CI: –0.18, 0.20), indicating no significant difference in AKI risk between the two drugs. The confidence interval crosses zero, and the point estimate is small.

Only one comparison between Furosemide and Prochlorperazine passed the criteria. With binary input and the S-Learner, the estimated CATE was 0.03 (95% CI: –0.22, 0.27), which is not statistically significant and indicates no clear direction of effect.

Several models were evaluated for the Furosemide with Lactulose as the negative drug. The most notable finding came from the Binary input with the X-Learner, which produced a CATE of –0.38 (95% CI: –0.65, –0.11), suggesting a statistically significant reduction in AKI risk associated with Furosemide relative to Lactulose. This is the only comparison where the CI does not cross zero.

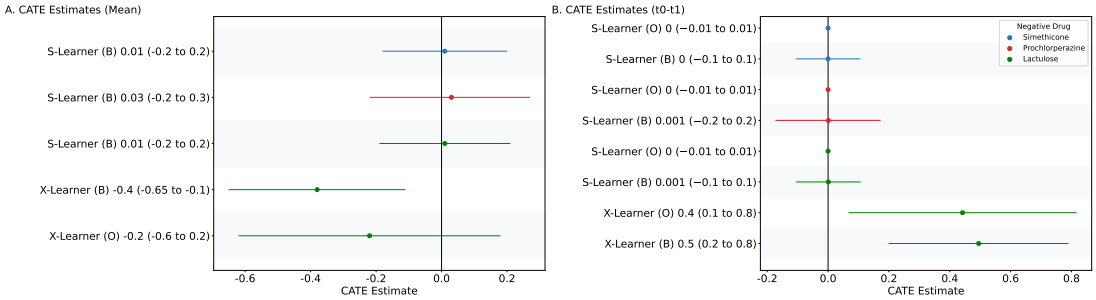

Figure 27: CATE Estimates (with 95% CI) for Furosemide

In furosemide (laboratory as first and last value) S-learner and X-learner calculated the most valid results with 6 and 2 CATEs out of 18, respectively. CATE estimates from comparisons with Simethicone were negligible across all models and encodings. Both the S-Learner with original input and the S-Learner with binary input yielded near-zero CATEs (e.g., 0.000 [95% CI: −0.007, 0.007]), indicating no meaningful difference in AKI risk between Furosemide and Simethicone. These results suggest that Simethicone may be a suitable negative control in similar causal inference frameworks.

As with Simethicone, Prochlorperazine produced CATEs that were nearly zero across all models and input formats, with S-Learner estimates such as 0.001 (95% CI: −0.173, 0.175). This suggests no significant differential effect on AKI risk when comparing Furosemide with Prochlorperazine.

Lactulose served as an informative comparator in this case study. The X-Learner yielded consistently positive and statistically significant CATEs across both original and binary input encodings, with values of 0.442 (95% CI: 0.067, 0.817) and 0.495 (95% CI: 0.199, 0.790), respectively. These findings strongly suggest that Furosemide is associated with a higher risk of AKI compared to Lactulose. In contrast, the S-Learner produced near-zero CATEs (close to 0 with very narrow CIs), which may reflect model insensitivity or lack of robustness in the low signal-to-noise scenario.

These findings indicate that Furosemide is associated with a significantly higher risk of AKI (Figure 27).

J.0.6. PANTOPRAZOLE

In pantoprazole (laboratory as mean values), there are 6 valid CATEs of S-learner and one of X-learner out of 18 (Figure 28). High CATE estimates are revealed in the comparison of Pantoprazole with Simethicone. The S-Learner suggests a modestly positive association, especially under binary encoding. However, the wide confidence intervals that include zero and range close to the 0.85 CI width threshold warrant caution in interpretation. The comparison with Simethicone shows slightly

high CATE estimates for Pantoprazole. However, the estimates remain non-significant and uncertain due to wide confidence intervals.

While the CATEs are slightly positive in the Pantoprazole with Prochlorperazine dataset, the confidence intervals remain wide and include zero in both original and binary encodings. There is no statistically significant indication of elevated AKI risk. Although the point estimates are slightly elevated, they remain statistically insignificant. No causal evidence supports an increased AKI risk with Pantoprazole relative to Prochlorperazine.

Pantoprazole with Lactulose, across both the original and binary data transformations using S-Learner, no evidence of increased or decreased risk of AKI associated with Pantoprazole use was observed. CATEs remained close to zero, and the confidence intervals were narrow and centered around zero. The estimated effect of Pantoprazole versus Lactulose on AKI is negligible. The small CATE values and tight confidence intervals that include zero suggest no significant association.

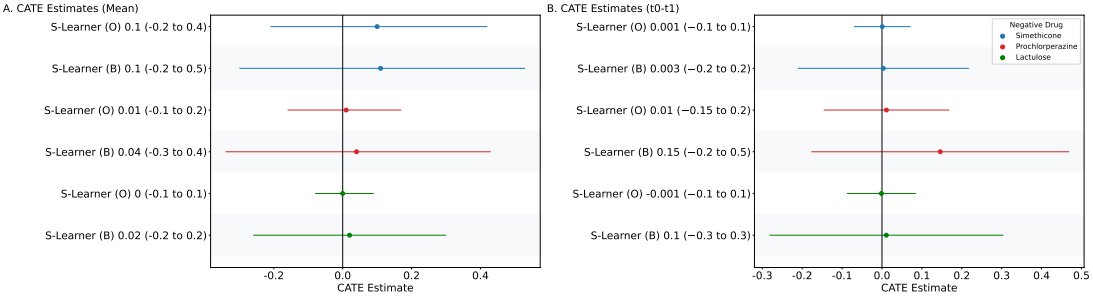

Figure 28: CATE Estimates (with 95% CI) for Pantoprazole

In the datasets where laboratory tests are first and last value of pantoprazole, there are 6 valid CATEs of S-learner and one of X-learner out of 18. Figure 28 summarizes the results of Pantoprazole versus Simethicone. Across all tested input representations (Original and Binary), the estimated CATEs were very close to zero with narrow confidence intervals, indicating no evidence of increased AKI risk from Pantoprazole when contrasted against Simethicone. These stable estimates suggest that both agents may be used in non-overlapping or similarly low-risk populations in terms of AKI, and that confounding was well controlled.

When comparing Pantoprazole to Prochlorperazine, the estimated CATEs were slightly higher than in the previous contrasts, particularly under the binary input model, where the CATE reached 0.146, although still with a wide confidence interval that included zero. This may suggest a modest trend toward increased AKI risk, though it did not reach statistical significance.

The comparison with Lactulose yielded similarly negligible effects. Across both original and binary inputs, the CATEs remained close to zero, and the confidence intervals were narrow enough to rule out any substantial increase in AKI risk. These findings suggest Pantoprazole does not confer increased risk relative to Lactulose, and any observed differences are likely due to random variability rather than a true causal relationship.

Overall, these findings do not provide evidence to support a causal relationship between Pantoprazole and AKI within the examined population and methodological boundaries.

J.0.7. OMEPRAZOLE

S-learner only produced valid CATEs (6/18). In Omeprazole (laboratory tests as mean values) with Simethicone dataset, CATE estimates are low, with narrow confidence intervals that include zero for both the original and binary input data types. This indicates no strong evidence of an association between Omeprazole and AKI compared to Simethicone.

The results of the CATE for Omeprazole versus Prochlorperazine. Similar to the Simethicone comparison, the CATE values remain close to zero across both the original and binary encodings. Confidence intervals remain narrow and include zero, indicating no significant difference between Omeprazole and Prochlorperazine in terms of AKI risk.

Omeprazole with Lactulose as the negative drug for both original and binary encodings, the CATE is close to zero with narrow confidence intervals that include zero, suggesting no significant association between Omeprazole and AKI when compared to Lactulose.

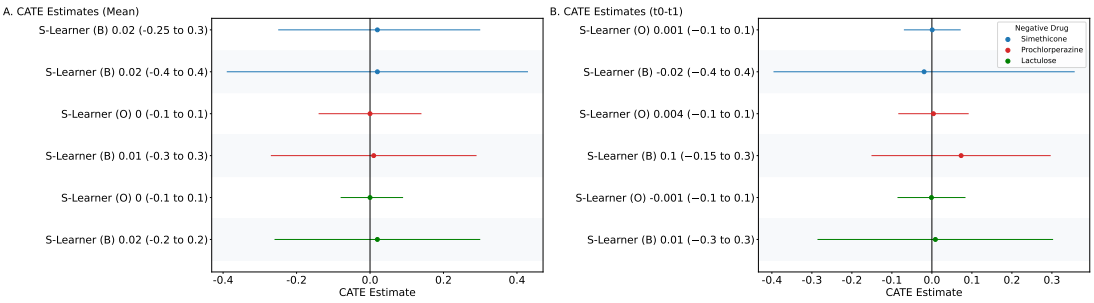

Figure 29: CATE Estimates (with 95% CI) for Omeprazole

S-learner and T-learner only produced valid CATEs in laboratory test representation as first and last value, 6 and one out of 18, respectively. When Omeprazole was compared with Simethicone, the CATEs were close to zero across both original and binary inputs. The original input yielded a CATE of 0.0008 (95% CI: -0.092, 0.094), and the binary input yielded -0.0195 (95% CI: -0.396, 0.357). These results indicate no meaningful effect on AKI risk.

In the comparison with Prochlorperazine, the original input yielded a CATE of 0.0038 (95% CI: -0.084, 0.092), while the binary input showed 0.0731 (95% CI: -0.151, 0.297). Though slightly higher than in other comparisons, the confidence intervals still cross zero and remain within acceptable bounds, indicating a non-significant association.

All CATEs were small and had confidence intervals that included zero. These consistent findings suggest that Omeprazole does not confer additional AKI risk relative to the chosen control drugs, even under varied meta-learning and input configurations (Figure 29).

J.0.8. ALLOPURINOL

The valid CATEs in allopurinol where laboratory tests are mean values (Figure 30), S-learner has the most results (11/18) and DML and X-learner have both a small number of accepted CATEs (3/18). CATEs from comparisons between Allopurinol and Simethicone. Across both nephrologist and literature reviews, CATE values are negative but small, and the confidence intervals include zero in all cases. The results indicate no statistically significant evidence of Allopurinol increasing or decreasing AKI risk compared to Simethicone.

CATE estimates for Allopurinol with Prochlorperazine are small and non-significant across both nephrologist and literature reviews. All CIs are tight and include zero, indicating stability in the estimates and no evidence of elevated AKI risk associated with Allopurinol.

All CML models and input combinations for Allopurinol compared to Lactulose show negative but small CATEs, with confidence intervals that include zero and remain within acceptable width limits. Across all modeling approaches, there is no statistically significant association observed between Allopurinol and AKI in comparison with Lactulose. The narrow CIs and consistent null findings support the absence of a strong effect. Finally, there is no valid CATE value based on the literature review DAG.

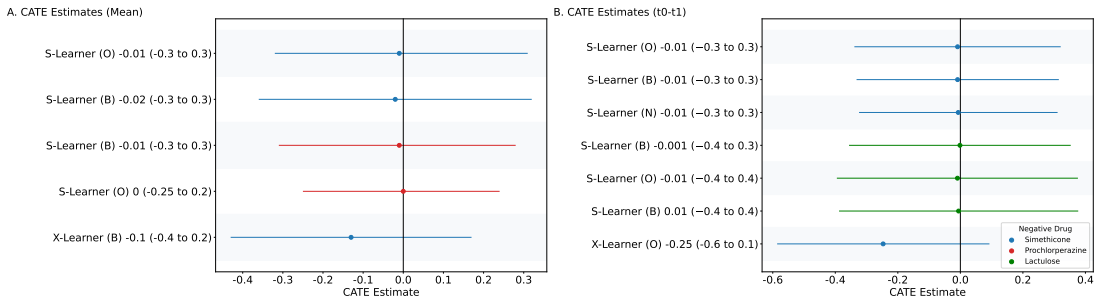

Figure 30: CATE Estimates (with 95% CI) for Allopurinol

In Allopurinol (Figure 30) where laboratory tests are first and last value, S-learner has the most results (7/18) and DML and X-learner have both a small number of accepted CATEs, 3 and one, respectively. The comparison with Simethicone yielded one notably informative result: using the X-Learner with original input, the CATE was -0.247 (95% CI: -0.586, 0.093). This suggests a potential protective effect of Allopurinol against AKI, although the CI slightly overlaps zero. While this estimate does not meet the threshold for definitive inference, the direction and magnitude suggest a signal worth further exploration. Other estimates from the S-Learner under binary, original, and normalized inputs remained close to zero, indicating no strong or consistent signal.

All comparisons between Allopurinol and Lactulose consistently yielded CATEs near zero, with narrow confidence intervals. For example, the original input using S-Learner returned a CATE of -0.0013 (95% CI: -0.356, 0.353). These results suggest no meaningful association between Allopurinol and AKI risk when compared to Lactulose, a pharmacologically distinct agent.

No results met the inclusion criteria (CI width < 0.85) for comparisons with Prochlorperazine. Therefore, this negative drug is excluded from formal reporting in this section. This likely reflects either insufficient sample balance or variance inflation, making inference unreliable in this comparison.

Taken together, the results suggest that Allopurinol does not exhibit a consistently increased risk of AKI across control comparisons. The strongest (though non-significant) signal of potential benefit emerged in comparison with Simethicone under the X-Learner, hinting at a possible reduction in AKI risk. However, the lack of consistent and significant findings across learners and controls supports a neutral causal relationship between Allopurinol use and AKI, under the current observational design.

J.0.9. AVERAGE CONDITIONAL AVERAGE TREATMENT EFFECT VALUES

To evaluate the CML methodologies in each suspect drug, we averaged the valid CATE estimations (CI < 0.85) for each CML model and the total CATE average of all the CML models. Figure 31 presents the mean and the standard deviation (SD) of every CML method for the data preprocessing method that includes the most valid CATEs in each drug (O, B, N). In vancomycin, pantoprazole and omeprazole, the total mean, which presents drug-induced AKI (positive effect) and SD were the same with S-learner as it is the only CML method that produced valid CATEs.

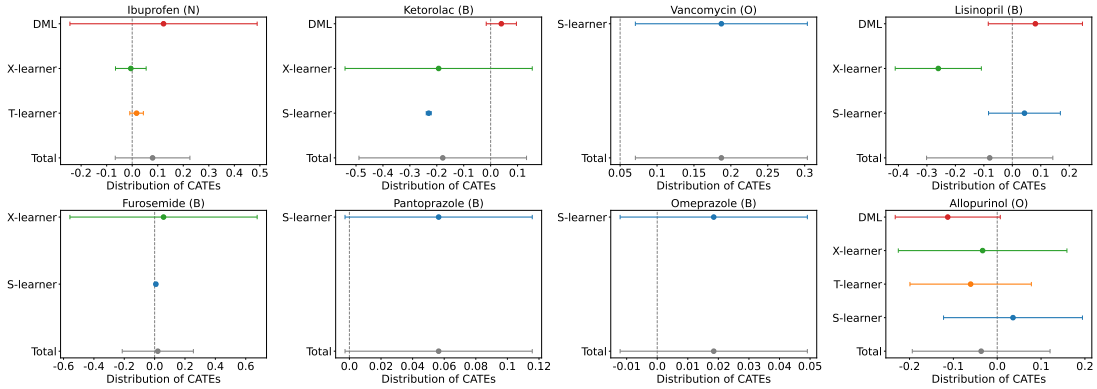

Figure 31: Average CATE Estimates (mean, standard deviation) per suspect drug and CML model

In Ibuprofen DML, T-learner and total average present positive CATEs with 0.12 (SD:0.4), 0.02 (SD: 0.03) and 0.08 (SD: 0.15), respectively. Only X-learner presents a slightly negative effect with -0.005 CATE (SD: 0.06). In contrast to the total positive average effect of ibuprofen, ketorolac's total effect is negative with CATE value of -0.18 (SD: 0.31), affected mainly from the negative average of X and S-learners with CATEs of -0.2 (SD: 0.34) AND -0.23 (SD: 0.01), respectively. DML produced a positive effect which indicates the adverse effect of AKI of ketorolac with CATE equals 0.04 (SD: 0.06). The negative total of Lisinopril drug, -0.08 (SD: 0.22), affected mainly from the strong negative CATE of X-leaner, -0.26 (SD: 0.15). On the other hand, S-learner and DML predict a the AKI adverse effect of the drug with CATE 0.04 (SD: 0.1) and 0.08 (SD: 0.16), respectively. The total average of furosemide of 0.02 (SD: 0.2) reflects the positive CATEs of the S-learner, 0.01 (SD: 0.01), and X-learner 0.06 (SD: 0.6). Finally, Allopurinol has total average -0.04 (SD: 0.2) which reflects the CATEs of the majority of CML architectures where T-learner is -0.06 (SD: 0.1), X-learner is -0.03 (SD: 0.2) and DML is -0.1 (SD: 0.1). The possibility of AKI adverse effect of Allopurinol is mirroring in S-learner CATE of 0.04 (SD: 0.2).

Appendix K. Heterogeneous Treatment Effect

The CML methodologies are selected for the calculation of HTEs based on the robustness of the width of CI (< 0.85) of CATE, their significance (preferred CIs that not contain 0) and their alignment with the literature such as Ketorolac which mainly produced negative CATE results but in this section we selected the X-learner CML model, mean laboratory data, Lactoluse negative drug and literature DAG architecture that produce positive CATE result of 0.2 (-0.02, 0.4). As the drugs are

not compared because of the diverse of the approaches, the HTE results are discussed as separately case studies (Table 16).

Suspect Drug (lab tests)	Negative drug	CML (input)	DAG	CATE (CI)
Furosemide (t0-t1)	Lactulose	X-Learner (B)	Literature/ Nephrologist	0.5 (0.2, 0.8)
Ibuprofen (mean)	Prochlorperazine	DML-Learner (N)	Nephrologist	0.3 (0.2, 0.4)
Vancomycin (t0-t1)	Lactulose	S-Learner (O)	Literature/ Nephrologist	0.06 (−0.3, 0.4)
Pantoprazole (t0-t1)	Simethicone	S-Learner (O)	Literature/ Nephrologist	0.001 (−0.07, 0.07)
Ketorolac (mean)	Lactulose	X-Learner (N)	Literature	0.2 (−0.02, 0.4)
Omeprazole (mean)	Lactulose	S-Learner (O)	Literature/ Nephrologist	0.02 (−0.3, 0.3)
Allopurinol (mean)	Simethicone	S-Learner (O)	Nephrologist	−0.05 (−0.5, 0.4)
Lisinopril (t0-t1)	Prochlorperazine	X-Learner (N)	Literature	0.45 (0.06, 0.85)

Table 16: Heterogeneous treatment effect models and characteristics of datasets for each suspect drug

K.0.1. IBUPROFEN

Demographics Analyzing the distribution of age and weight for patients with no effect and patients with adverse effect, we conclude that patients with ibuprofen-induced AKI are significantly older and heavier (Figure 32).

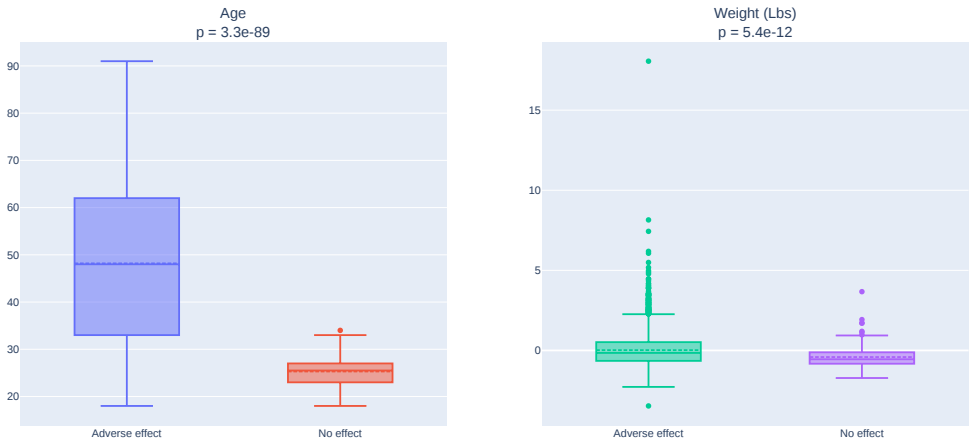

Figure 32: Distribution of HTEs according to age and weight in Ibuprofen

Age The HTE mean for ibuprofen increases from 0.20 in young adults (18–39) to a peak of 0.42 in middle-aged adults (40–59), then slightly decreases in older groups (0.34 in 60–79, 0.33 in ≥80). The standard deviation (SD) decreases with age, suggesting more consistent effects in older adults.

Gender Males (HTE mean 0.368) show a higher average effect than females (0.291), with lower variability (SD 0.064 vs 0.122).

Demographic	Category	N	HTE μ (σ)
Gender	Female	3435	0.3 ($\pm$ 0.1)
	Male	1312	0.4 ($\pm$ 0.1)
Age	18–39	1818	0.2 ($\pm$ 0.1)
	40–59	1515	0.42 ($\pm$ 0.05)
	60–79	1254	0.34 ($\pm$ 0.01)
	$\geq$80	160	0.33 ($\pm$ 0)
Weight	$<$60 kg	1	0.355
	60–79 kg	9	0.4 ($\pm$ 0.05)
	80–99 kg	59	0.3 ($\pm$ 0.1)
	$\geq$100 kg	4678	0.3 ($\pm$ 0.1)

Table 17: Heterogeneous Treatment Effect (HTE) by Demographic Category in Ibuprofen

Weight The majority of patients taking ibuprofen fall into the obese/very overweight category, weighing $\geq$ 100 kg, with an average HTE of 0.311 (SD=0.11; N=4678). A smaller group of patients is categorized as overweight, weighing between 80–99 kg, and exhibits an average HTE of 0.34 (SD=0.08; N=59). The group of normal weight patients, weighing 60–79 kg, is notably small, with an HTE average of 0.4 (SD=0.05; N=9). In the underweight category, weighing less than 60 kg, there is only one patient with an HTE of 0.355. Due to the limited number of patients in the normal weight and underweight categories, it is not possible to draw any definitive conclusions.

Table 17 describes the mean HTE in the different values of each demographic category. Overall, the highest mean HTE values, concerning an adequate number of patients, are detected to middle-aged (40-59), males and overweight patients. The standard deviations are small, reflecting robust results in all demographic categories.

Laboratory tests Laboratory results indicate that patients with elevated levels of glucose, creatinine, blood urea nitrogen (BUN), and bicarbonate, as well as reduced levels of chloride and anion gap, are significantly more likely to develop ibuprofen-induced AKI (p ¡ 0.01). Moreover, patients with higher potassium levels exhibit a statistically significant difference (p ¡ 0.05) when comparing those with positive HTEs to those with HTEs in the range of -0.1 to 0.1 (Figure 33).

Table 18 presents the detailed analysis based on the normal and abnormal values of in each laboratory test.

Glucose In patients with normal fasting glucose ($<$ 100 mg/dL), Ibuprofen yielded a mild positive HTE mean of 0.263 (SD = 0.13; N = 1663). In the pre-diabetic (100–125 mg/dL), the mean HTE is slightly increased to 0.33 (SD = 0.1; N = 1952). In diabetic patients ($\geq$ 126 mg/dL), HTE mean is the highest with 0.356 (SD=0.07; N = 1132).

Sodium Among hyponatremic patients ($<$ 135 mEq/L), Ibuprofen was associated with a MODERATE HTE increase of 0.342 (SD = 0.1; N = 346). More of the patients belong in the category of normonatremic (135–145 mEq/L) with a mild HTE mean to 0.31 (SD = 0.11; N = 4388). An insignificant number of patients in terms of evidence belong to the hypernatremic category, with HTE mean 0.366 (SD = 0.1; N = 13).

Creatinine Most of ibuprofen patients are categorized in the normal creatinine group ($<$ 1.2 mg/dL), exhibited a mild HTE of 0.3 (SD = 0.11; N = 4414), whereas in the mildly elevated (1.2–1.9

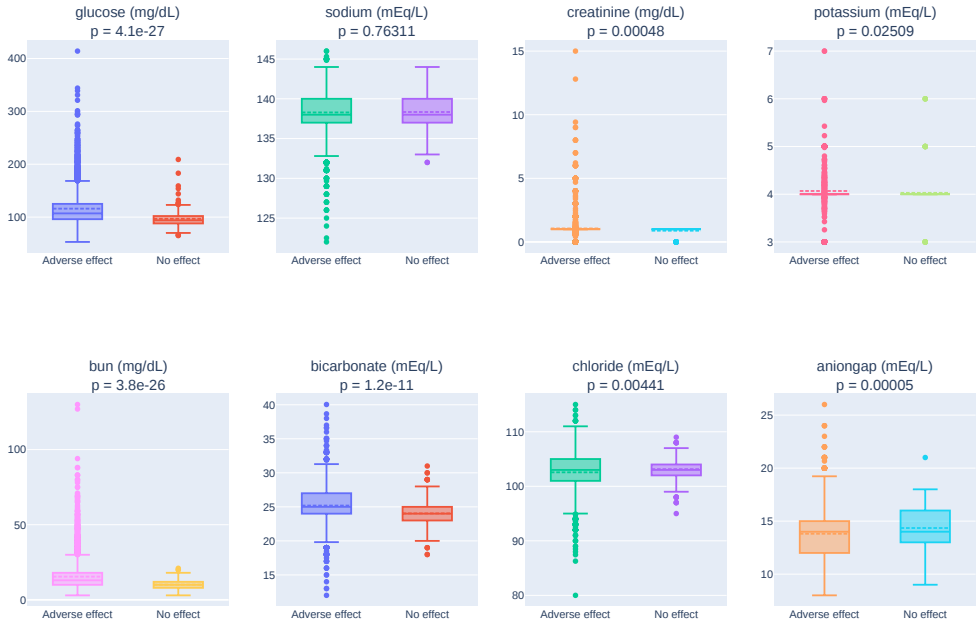

Figure 33: Distribution of HTEs according to different laboratory tests in Ibuprofen.

mg/dL) the number of patients is relatively small with moderate CATE 0.411 (SD=0.04; N = 26). At severely elevated ($\geq$ 2.0 mg/dL) strata, HTE mean was 0.371 (SD = 0.05; N=307).

Potassium Ibuprofen-treated patients with hypokalemia ($<$ 3.5 mEq/L) had an HTE of 0.27 (SD = 0.121; N = 117), while those with normal potassium values (7–20 mEq/L) which includes the majority of the patients, showed a slightly higher HTE mean of 0.312 (SD = 0.11, N = 4232). In patients with hyperkalaemia ($>$ 5.0 mEq/L), the HTE mean reaches the highest values of 0.32 (SD = 0.1, N = 398)

BUN Patients with BUN values below normal ($<$ 7 mg/dL), the HTE mean was mild 0.235 (SD = 0.12; N = 219). The majority of patients belong to the normal category (7–20 mg/dL) with HTE mean 0.3 (SD = 0.12; N = 3661) and in the elevated BUN levels ($>$ 20 mg/dL) group, HTE mean was highest with 0.361 (SD = 0.05; N = 867).

Bicarbonate Patients with metabolic acidosis ($<$ 22 mEq/L) administered Ibuprofen demonstrated an HTE of 0.27 (SD = 0.12; N = 375) while in patients with normal values (22–29 mEq/L) the HTE mean is slightly increased to 0.312 (SD = 0.11; N = 3958). Finally, in patients with metabolic alkalosis ($>$ 29 mEq/L) the HTE mean has the highest value of 0.35 (SD = 0.1; N = 414).

Chloride In the hypochloremic cohort ($<$98 mEq/L), HTE was moderate to 0.362 (SD = 0.118; N = 272) while in hyperchloremic patients ($>$106 mEq/L) HTE mean was lower, 0.313 (SD = 0.11; N = 697). The normal chloride group of patients presents the lower HTE mean value compared to the other two categories, with 0.31 (SD = 0.11; N = 3778)

Anion Gap Patients with a low anion gap (¡ 8 mEq/L) are not represented in our dataset. For the patients with normal values (8–16 mEq/L), the mean HTE is 0.32 (SD = 0.11; N = 3855). For

Laboratory Test	Category	N	HTE μ (σ)
Glucose	<100 mg/dL	1663	0.263 ($\pm$ 0.1)
	100–125 mg/dL	1952	0.33 ($\pm$ 0.11)
	$\geq$ 126 mg/dL	1132	0.4 ($\pm$ 0.1)
Sodium	<135 mEq/L	346	0.342 ($\pm$ 0.1)
	135–145 mEq/L	4388	0.31 ($\pm$ 0.115)
	> 145 mEq/L	13	0.366 ($\pm$ 0.1)
Creatinine	<1.2 mg/dL	4414	0.31 ($\pm$ 0.1)
	1.2–1.9 mg/dL	26	0.4 ($\pm$ 0.04)
	$\geq$2.0 mg/dL	307	0.4($\pm$ 0.05)
Potassium	<3.5 mEq/L	117	0.3 ($\pm$ 0.1)
	3.5-5.0 mEq/L	4232	0.3 ($\pm$ 0.1)
	>5.0 mEq/L	398	0.3 ($\pm$ 0.1)
BUN	<7 mg/dL	219	0.2 ($\pm$ 0.1)
	7–20 mg/dL	3661	0.3 ($\pm$ 0.1)
	¿20 mg/dL	867	0.361 ($\pm$ 0.05)
Bicarbonate	<22 mEq/L	375	0.3 ($\pm$ 0.1)
	22–29 mEq/L	3958	0.3 ($\pm$ 0.1)
	¿29 mEq/L	414	0.3 ($\pm$ 0.1)
Chloride	<98 mEq/L	272	0.362 ($\pm$ 0.1)
	98–106 mEq/L	3778	0.362 ($\pm$ 0.1)
	>106 mEq/L	697	0.3 ($\pm$ 0.1)
Anion Gap	8–16 mEq/L	3855	0.3 ($\pm$ 0.1)
	>16 mEq/L	892	0.3 ($\pm$ 0.1)

Table 18: Ibuprofen: Heterogeneous Treatment Effect (HTE) by Laboratory Category

the elevated (¿ 16 mEq/L) group, the HTE mean is lower than the normal, with 0.28 (SD = 0.12; N = 892).

Overall, patients with high mean HTE according to laboratory values are characterized from elevated levels of glucose ($\geq$ 126 mg/dL), and BUN (>20 mg/dL), low values of sodium (< 135 mEq/L) and normal values of chloride (98–106 mEq/L) in patients (Table 18).

Blood pressure In figure 34, it is obvious that patients with AKI adverse effect have higher blood pressure than those with no AKI effect, with statistical significance ($p < 0.01$).

In patients with normal systolic blood pressure (<120 mmHg), ibuprofen shows a moderate positive HTE mean of 0.312 (SD 0.114, N=4747). This suggests a measurable heterogeneous treatment effect in normotensive individuals. No data are available for hypertensive categories, so conclusions are limited to those with normal baseline systolic pressure. For diastolic blood pressure ¡80 mmHg, the HTE mean is 0.375 (SD 0.04, N=2), suggesting a moderate influence among normotensive subjects, though the limited sample size is insufficient for conclusive evidence. A small cohort falls into the initial stage of hypertension (80-89 mmHg), where the mean HTE is similarly moderate at 0.35 (SD 0.1, N=8). The majority of patients taking ibuprofen belong to the second stage of hypertension, with diastolic pressure $\geq$ 90 mmHg, exhibiting a moderate HTE mean of 0.312 (SD 0.11, N=4737).

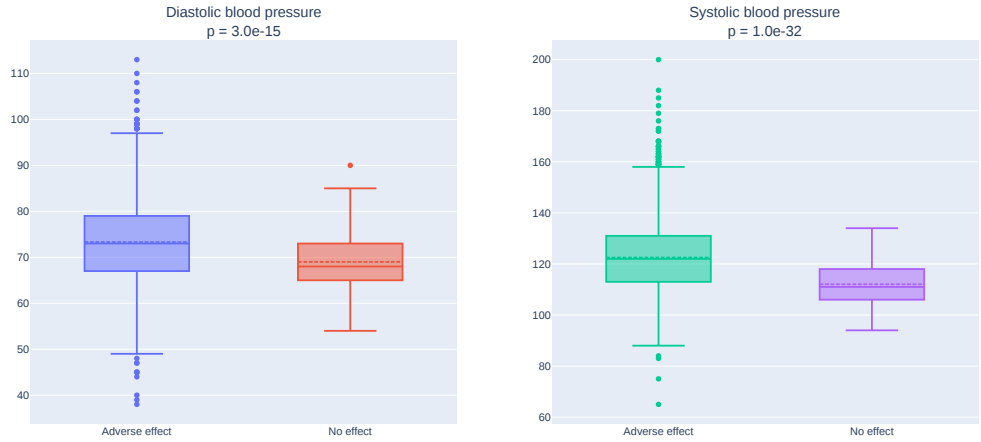

Figure 34: Distribution of HTEs according to diastolic and systolic blood pressure in Ibuprofen

K.0.2. KETOROLAC

Demographics Similar to the trend observed with ibuprofen, older patients are more likely to experience ketorolac-induced AKI, whereas younger patients tend to have HTEs between -0.1 and 0.1. This age difference is statistically significant (p ¡ 0.01). In contrast, there is no statistically significant difference in weight between patients who experienced an adverse effect and those who did not (Figure 35).

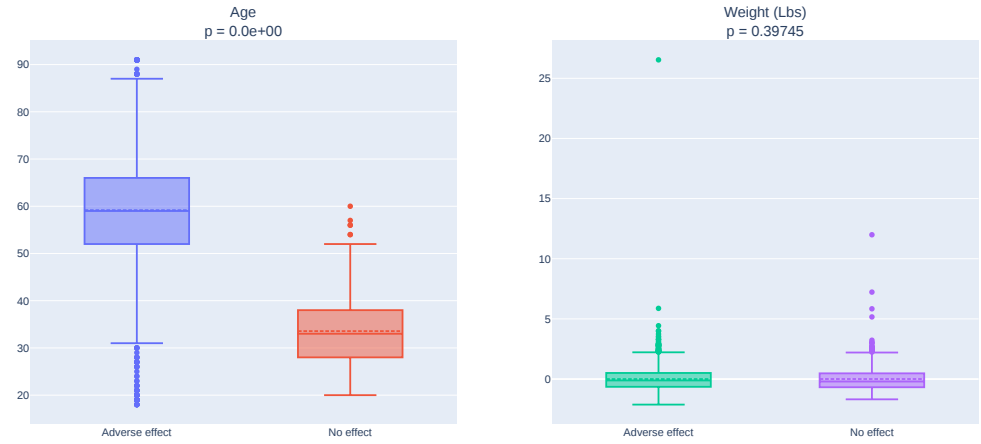

Figure 35: Distribution of HTEs according to age and weight in Ketorolac

Age HTE mean rises steadily with age: 0.06 (18–39), 0.2 (40–59), 0.39 (60–79), and 0.50 ($\geq$80), with SD also increasing.

Gender Males (0.287) show higher HTE mean than females (0.211), with slightly lower SD, mirroring ibuprofen trends (Table 19).

Weight Underweight/low-weight individuals ($<$ 60 kg) are not represented in our dataset, which is a notable gap. The majority of patients fall into the obese/very overweight category, weighing $\geq$ 100 kg, with an average HTE of 0.243 (SD=0.163; N=4041). A smaller group of patients is categorized as overweight, weighing between 80–99 kg, and exhibits an average HTE of 0.288 (SD=0.2; N=54). The group of normal weight patients, weighing 60–79 kg has only one patient, with HTE 0.173. Due to the limited number of patients in the normal weight category, it is not possible to draw any definitive conclusions.

Table 19 summarizes the mean HTE in the demographic categories. The highest mean HTE values, concerning an adequate number of patients, are detected in the aged (60-79), males and overweight patients. The standard deviations are small, reflecting robust results in all demographic categories.

Demographic	Category	N	HTE μ (σ)
Gender	Female	2338	0.2 ($\pm$ 0.2)
	Male	1758	0.3 ($\pm$ 0.2)
Age	18–39	971	0.06 ($\pm$ 0.04)
	40–59	1585	0.2 ($\pm$ 0.1)
	60–79	1344	0.39 ($\pm$ 0.1)
	$\geq$80	196	0.5 ($\pm$ 0.1)
Weight	60–79 kg	1	0.2
	80–99 kg	54	0.3 ($\pm$ 0.2)
	$\geq$100 kg	4041	0.2 ($\pm$ 0.2)

Table 19: Heterogeneous Treatment Effect (HTE) by Demographic Category for Ketorolac

Laboratory tests In ketorolac patients with AKI as an adverse effect having significant higher level of glucose, creatinine, bun, potassium and bicarbonate and lower sodium, chloride and anion gap than those who do not have AKI adverse effect. The statistical significance of these differences is $p < 0.01$ (Figure 36). In contrast with ibuprofen, in ketorolac sodium between the two categories of patients has a statistical significance.

Glucose In patients with normal fasting glucose ($<$ 100 mg/dL), ketorolac patients yielded a low positive HTE mean of 0.164 (SD = 0.142; N = 1084). In the pre-diabetic (100–125 mg/dL), the mean HTE is increased to 0.242 (SD = 0.156; N = 1808). In diabetic patients ($\geq$ 126 mg/dL), HTE mean is the highest with 0.317 (SD=0.16; N = 1204).

Sodium Among hyponatremic patients ($<$ 135 mEq/L), ketorolac administration was associated with a mild HTE increase of 0.3 (SD = 0.15; N = 535). More of the patients belong in the category of normal values (135–145 mEq/L) with a low HTE mean to 0.236 (SD = 0.164; N = 3543). An insignificant number of patients in terms of evidence belong to the hypernatremic category, with HTE mean 0.4 (SD = 0.134; N = 18).

Creatinine Most of the patients are classified in the normal creatinine group ($<$ 1.2 mg/dL), exhibited a low HTE of 0.223 (SD = 0.16; N = 3432), whereas in the mildly elevated (1.2–1.9 mg/dL) group the number of patients is relatively small with mild CATE 0.346 (SD=0.154; N = 102). At severely elevated ($\geq$ 2.0 mg/dL) strata, HTE mean were the highest with 0.352 (SD = 0.142; N=562).

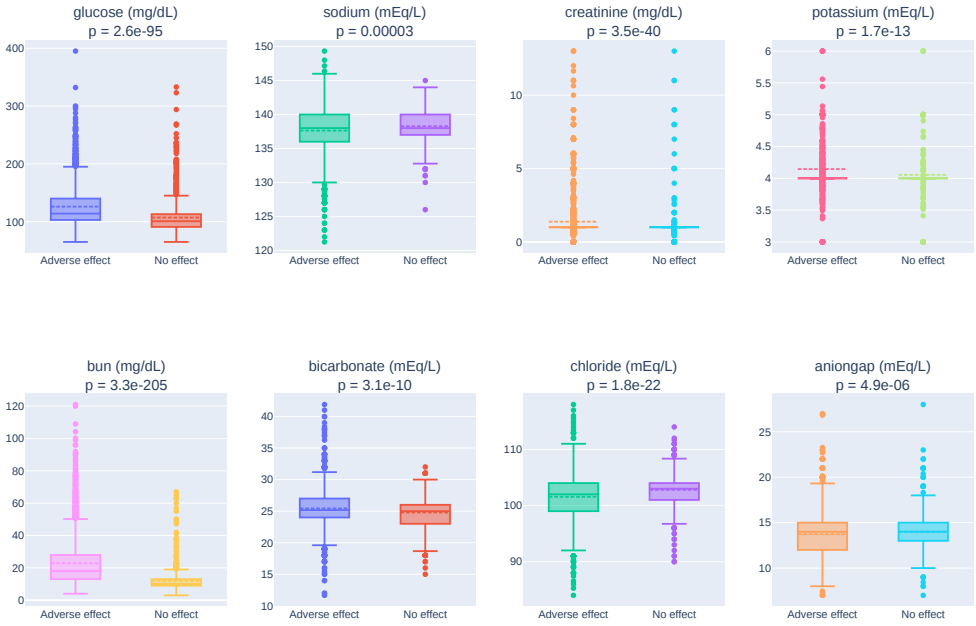

Figure 36: Distribution of HTEs according to different laboratory tests in Ketorolac

Potassium Patients with hypokalemia (< 3.5 mEq/L) had an HTE of 0.15 (SD = 0.14; N = 46), while those with normal potassium values (7–20 mEq/L), which includes the majority of the patients, showed a higher HTE mean of 0.24 (SD = 0.163, N = 3571). In patients with hyperkalaemia (> 5.0 mEq/L) the HTE mean reaches the highest values of 0.286 (SD = 0.156, N = 479)

BUN BUN values below normal (< 7 mg/dL) include patients with low HTE mean of 0.088 (SD = 0.07; N = 120). The majority of patients belong to the normal category (7–20 mg/dL) with mild HTE mean 0.2 (SD = 0.15; N = 2631) and in the elevated BUN levels (> 20 mg/dL) group, HTE mean was was highest with 0.35 (SD = 0.14; N = 1345).

Bicarbonate Patients with metabolic acidosis (< 22 mEq/L) demonstrated an HTE of 0.24 (SD = 0.15; N = 371) while in patients with metabolic alkalosis (> 29 mEq/L) the HTE mean has increased to 0.311 (SD = 0.166; N = 404). In the normal values (22–29 mEq/L) group that includes the majority of patients, the HTE mean is low, 0.236 (SD = 0.16; N = 3321). Finally, in patients with.

Chloride In the hypochloremic cohort (<98 mEq/L), HTE was moderate to 0.32 (SD = 0.155; N = 456) while in hyperchloremic (>106 mEq/L) and normal chloride (98–106 mEq/L) value patients present almost the same low HTE means with, 0.237 (SD = 0.16; N = 513) and 0.233 (SD = 0.162; N = 3127), respectively.

Anion Gap Patients with a low anion gap (¡ 8 mEq/L) are only 6 in our dataset with mean HTE 0.3 (SD = 0.2). For the normal value patients (8–16 mEq/L), the mean HTE is 0.244 (SD = 0.16; N = 3326). For the elevated (¿ 16 mEq/L) group, the HTE mean is almost the same as normal with 0.243 (SD = 0.174; N = 764).

Laboratory Test	Category	N	HTE μ (σ)
Glucose	<100 mg/dL	1084	0.164 ($\pm$ 0.1)
	100–125 mg/dL	1808	0.242 ($\pm$ 0.15)
	$\geq$ 126 mg/dL	1204	0.32 ($\pm$ 0.2)
Sodium	<135 mEq/L	535	0.3 ($\pm$ 0.1)
	135–145 mEq/L	3543	0.23 ($\pm$ 0.2)
	> 145 mEq/L	18	0.4 ($\pm$ 0.1)
Creatinine	<1.2 mg/dL	3432	0.223 ($\pm$ 0.2)
	1.2–1.9 mg/dL	102	0.346 ($\pm$ 0.15)
	$\geq$2.0 mg/dL	562	0.352 ($\pm$ 0.14)
Potassium	<3.5 mEq/L	46	0.15 ($\pm$ 0.14)
	3.5-5.0 mEq/L	3571	0.24 ($\pm$ 0.16)
	>5.0 mEq/L	479	0.3 ($\pm$ 0.16)
BUN	<7 mg/dL	120	0.1 ($\pm$ 0.07)
	7–20 mg/dL	2631	0.2 ($\pm$ 0.15)
	¿20 mg/dL	1345	0.35 ($\pm$ 0.14)
Bicarbonate	<22 mEq/L	371	0.24 ($\pm$ 0.1)
	22–29 mEq/L	3321	0.24 ($\pm$ 0.1)
	¿29 mEq/L	404	0.3 ($\pm$ 0.16)
Chloride	<98 mEq/L	456	0.32 ($\pm$ 0.15)
	98–106 mEq/L	3127	0.23 ($\pm$ 0.16)
	>106 mEq/L	513	0.24 ($\pm$ 0.16)
Anion Gap	< 8 mEq/L	6	0.3 ($\pm$ 0.2)
	8–16 mEq/L	3326	0.24 ($\pm$ 0.16)
	>16 mEq/L	764	0.24 ($\pm$ 0.17)

Table 20: Ketorolac: Heterogeneous Treatment Effect (HTE) by Laboratory Category

According to laboratory values presented in table 20, patients with high mean HTE are characterized from elevated levels of glucose ($\geq$ 126 mg/dL), BUN (>20 mg/dL), creatinine ($\geq$ 2.0 mg/dL), potassium ($>$ 5.0 mEq/L), and bicarbonate ($>$ 29 mEq/L), low values of sodium ($<$ 135 mEq/L), chloride ($<$ 98 mEq/L) in patients.

Blood pressure Diastolic blood pressure of patients with AKI adverse effect is significantly lower than patients without AKI adverse effect ($p < 0.01$). On the other hand, systolic blood pressure is significantly higher for patients with AKI adverse effects compared to those without (Figure 37).

In patients with normal blood pressure values, according to systolic blood pressure, the HTE mean is 0.244 (SD = 0.163, N = 4094), suggesting a low effect. In the category of elevated levels there are only 2 patients with a mean HTE of 0.08 (SD = 0.05). No data are available for hypertensive categories. In normal diastolic blood pressure levels (¡80 mmHg), there are only 2 patients with HTE mean of 0.438 (SD =0.2), which cannot give us strong evidence about the ketorolac-induced AKI effect in this category of patients. An also non non-informative small number of patients fall into the initial stage of hypertension (80-89 mmHg), where the mean HTE is 0.26 (SD = 0.2, N=11). The majority of patients belong to the second stage of hypertension, with diastolic pressure $\geq$ 90 mmHg, exhibiting a low HTE mean of 0.243 (SD = 0.163, N=4083).

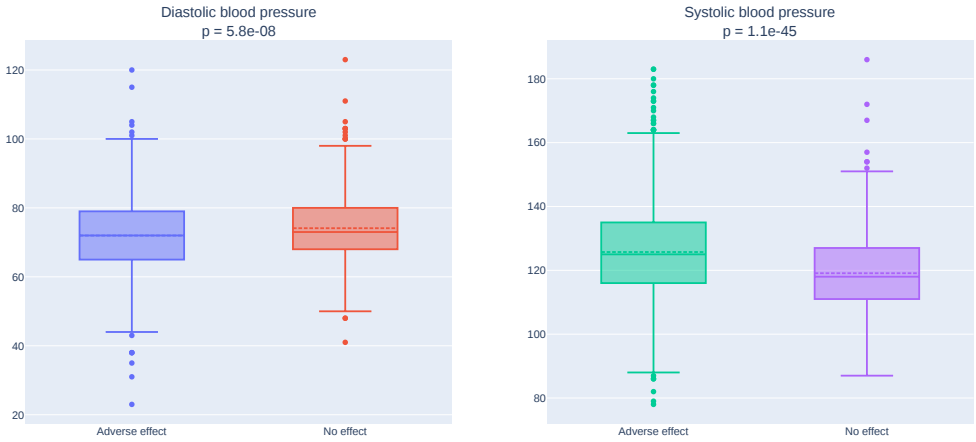

Figure 37: Distribution of HTEs according to blood pressure (diastolic, systolic) in Ketorolac

K.0.3. VANCOMYCIN

Demographics For both age and weight, the distribution of HTEs follows a similar pattern. Patients who experienced vancomycin-induced AKI tend to be younger and lighter than those with no effect, but older and heavier than those with a protective effect (Figure 38).

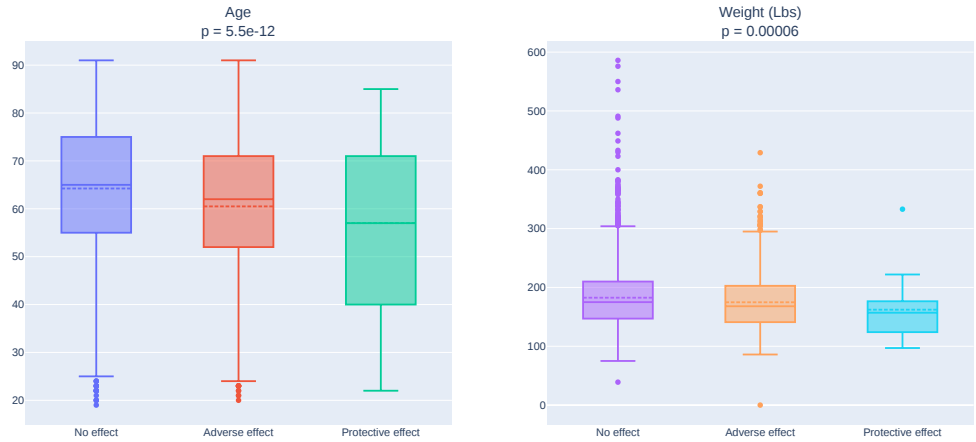

Figure 38: Distribution of HTEs according to age and weight in Vancomycin

Age The highest HTE mean is in young adults (0.39), decreasing with age to 0.15 in those $\geq$80. SD is high across all groups, indicating substantial variability in individual response.

Gender Females have a higher HTE mean (0.32) than males (0.226), but both have high SDs.

Demographic	Category	N	HTE μ (σ)
Gender	Female	1791	0.3 ($\pm$ 0.5)
	Male	2591	0.2 ($\pm$ 0.4)
Age	18–39	252	0.4 ($\pm$ 0.5)
	40–59	1318	0.3 ($\pm$ 0.5)
	60–79	2115	0.3 ($\pm$ 0.5)
	$\geq$80	697	0.15 ($\pm$ 0.4)
Weight	$<$60 kg	2	0.5 ($\pm$ 0.7)
	80–99 kg	62	0.3 ($\pm$ 0.5)
	$\geq$100 kg	4314	0.3 ($\pm$ 0.45)

Table 21: Heterogeneous Treatment Effect (HTE) by Demographic Category for Ketorolac

Weight Very limited data for low weight; more robust data are available only for overweight and obese categories, where HTE means mild to moderate (0.29–0.264).

Table 21 summarizes the mean HTE in the demographic categories. The highest mean HTE values, concerning an adequate number of patients, are detected to middle-aged and aged (40-79), females and overweight patients. The standard deviations are high, so the evidence is not robust and needs more investigation.

Laboratory tests In creatinine, potassium and anion gap, patients with vancomycin-induced AKI have significantly ($p < 0.01$) lower levels than patients with an effect close to zero (no effect) and higher than the patients with a negative or protective effect. The same pattern is followed in glucose, where the statistical significance is slightly bigger (0.03). In chloride, the pattern is completely the opposite (no effect has the lower level and a protective effect at the higher with an adverse effect between them) with statistical significance $p < 0.01$. Both bun and bicarbonate present similar results where patients with vancomycin-induced AKI have significant ($p < 0.01$) lower levels compare with the other two categories of patients.

Glucose Vancomycin had a minimal impact on fasting glucose: HTEs in the normal ($<$ 100 mg/dL), pre-diabetic (100–125 mg/dL) and diabetic ($\geq$ 126 mg/dL) groups were 0.282 (SD = 0.459; N = 1055), 0.273 (SD = 0.461; N = 1492) and 0.247 (SD = 0.434; N = 1835), respectively.

Sodium In hyponatremic patients ($<$ 135 mEq/L), HTE was 0.284 (SD = 0.459; N = 814), while in those with normal values (135–145 mEq/L) and hypernatremic ($>$ 145 mEq/L), HTEs were 0.263 (SD = 0.448; N = 3439) and 0.171 (SD = 0.398; N = 129).

Creatinine Normal-creatinine patients ($<$ 1.2 mg/dL) showed HTE of 0.312 (SD = 0.475; N = 2196), while mildly (1.2–1.9 mg/dL) and severely ($\geq$ 2.0 mg/dL) elevated groups had HTEs of 0.227 (SD = 0.423; N = 1248) and 0.203 (SD = 0.407; N = 938).

Potassium Hypokalemic ($<$ 3.5; N = 178) and hyperkalemic ($>$ 5.0; N = 289) HTEs were 0.393 (SD = 0.501) and 0.208 (SD = 0.415).

BUN Vancomycin-treated patients with low BUN ($<$ 7 mg/dL) had HTE mean 0.632 (SD = 0.506; N = 95), those with normal BUN (7–20 mg/dL) HTE mean is 0.319 (SD = 0.478; N = 1681), and elevated BUN ($>$ 20 mg/dL) HTE mean is 0.216 (SD = 0.417; N = 2606).

Bicarbonate In metabolic acidosis ($<$ 22 mEq/L) HTE was 0.299 (SD = 0.463; N = 934), with normal (22–29 mEq/L) and alkalotic ($>$ 29 mEq/L) groups at 0.263 (SD = 0.451; N = 2903) and 0.215 (SD = 0.411; N = 545), respectively.

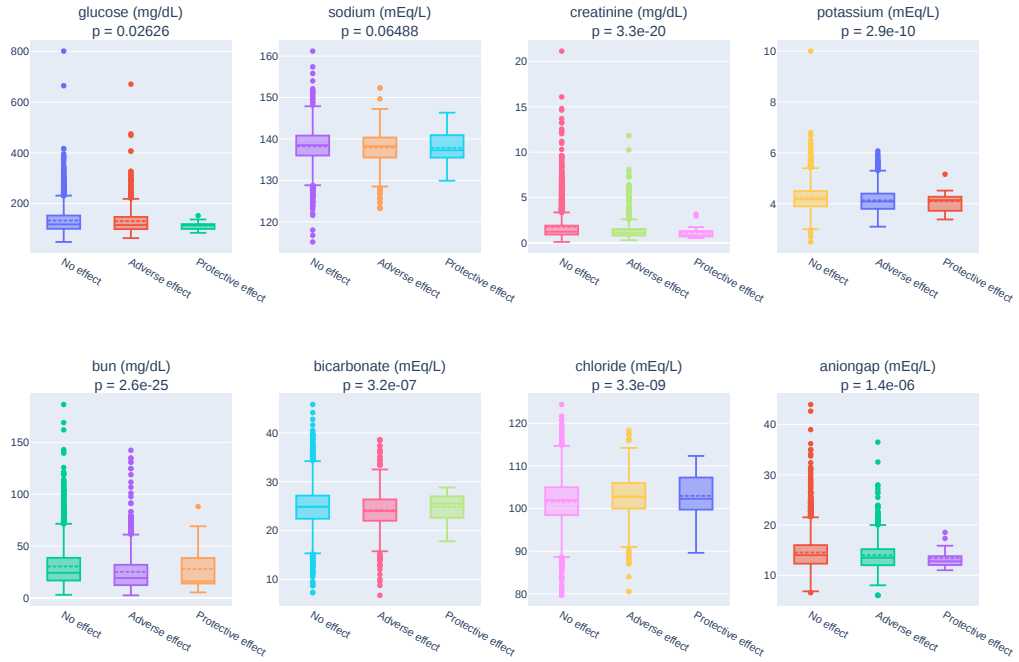

Figure 39: Distribution of HTEs according to different laboratory tests in Vancomycin

Chloride Hypochloremic ($<$ 98 mEq/L), normal chloride (98–106 mEq/L), and hyperchloremic ($>$ 106 mEq/L) HTEs were 0.205 (SD = 0.409; N = 898), 0.268 (SD = 0.450; N = 2556), and 0.314 (SD = 0.478; N = 928), respectively.

Anion Gap Normal anion gap (8–16 mEq/L), low anion gap ($<$ 8 mEq/L), and elevated anion gap ($>$ 16 mEq/L) showed HTEs of 0.281 (SD = 0.459; N = 3288), 0.235 (SD = 0.437; N = 17), and 0.215 (SD = 0.416; N = 1077).

Table 22 summarizes the different laboratory values categorization. Patients with high mean HTE are characterized from low levels of creatinine ($<$ 1.2 mg/dL) and normal values of anion gap (8–16 mEq/L).

Blood pressure Only diastolic blood pressure presents a significant ($p<$ 0.01) higher level for patients with AKI adverse effect compared those with no effect and lower compared those with protective effect.

The HTE mean in patients with normal values is 0.265 (SD 0.449, N=4381), with no data for hypertensive categories. For diastolic blood pressure $<$80 mmHg, the HTE mean is 0.25 (SD 0.500, N=4), but this is based on a very small sample and should be interpreted with caution. In stage 1 hypertension (80–89 mmHg), the HTE mean drops to 0.071 (SD 0.267, N=14), and in stage 2 hypertension ($\geq$ 90 mmHg), it is 0.265 (SD 0.450, N=4364), closely mirroring the systolic findings. The consistency of the HTE mean in stage 2 hypertension and normal groups.

Laboratory Test	Category	N	HTE μ (σ)
Glucose	<100 mg/dL	1055	0.3 ($\pm$ 0.5)
	100–125 mg/dL	1492	0.3 ($\pm$ 0.5)
	$\geq$126 mg/dL	1835	0.25 ($\pm$ 0.4)
Sodium	<135 mEq/L	814	0.3 ($\pm$ 0.5)
	135–145 mEq/L	3439	0.3 ($\pm$ 0.45)
	>145 mEq/L	129	0.2 ($\pm$ 0.4)
Creatinine	<1.2 mg/dL	2196	0.3 ($\pm$ 0.5)
	1.2–1.9 mg/dL	1248	0.2 ($\pm$ 0.4)
	$\geq$2.0 mg/dL	161	0.2 ($\pm$ 0.4)
Potassium	<3.5 mEq/L	178	0.4 ($\pm$ 0.5)
	3.5–5.0 mEq/L	3915	0.26 ($\pm$ 0.45)
	>5.0 mEq/L	289	0.2 ($\pm$ 0.4)
BUN	<7 mg/dL	95	0.6 ($\pm$ 0.5)
	7–20 mg/dL	1681	0.3 ($\pm$ 0.5)
	>20 mg/dL	2606	0.2 ($\pm$ 0.4)
Bicarbonate	<22 mEq/L	934	0.3 ($\pm$ 0.5)
	22–29 mEq/L	2903	0.3 ($\pm$ 0.45)
	>29 mEq/L	545	0.2 ($\pm$ 0.4)
Chloride	<98 mEq/L	898	0.2 ($\pm$ 0.4)
	98–106 mEq/L	2556	0.3 ($\pm$ 0.45)
	>106 mEq/L	928	0.3 ($\pm$ 0.5)
Anion Gap	<8 mEq/L	17	0.2 ($\pm$ 0.5)
	8–16 mEq/L	3288	0.3 ($\pm$ 0.4)
	>16 mEq/L	1077	0.2 ($\pm$ 0.4)

Table 22: Vancomycin: Heterogeneous Treatment Effect (HTE) by Laboratory Category

K.0.4. LISINOPRIL

Demographics Patients with AKI adverse effect are significant ($p < 0.01$) younger and heavier compare to patients with no or protective effect (Figure 41).

Age HTE means are high in all but the oldest group: 0.58 (18–39), 0.62 (40–59), 0.55 (60–79), but -0.07 in $\geq$80, suggesting reduced or potentially adverse effect in the oldest adults. This may reflect age-related changes in the renin-angiotensin system and higher prevalence of comorbidities in the elderly.

Gender Similar HTE means for females (0.467) and males (0.454), with high SDs, indicating consistent but variable responses across sexes.

Weight Underweight/low-weight individuals ($<$ 60 kg) are not represented in our dataset, creating an important gap. The majority of patients fall into the obese/very overweight category, weighing $\geq$ 100 kg, with a moderate HTE of 0.46 (SD = 0.352; N = 3717). A relative small group of patients is categorized as overweight, weighing between 80–99 kg, and exhibits an average HTE of 0.287 (SD = 0.366; N=45). The group of normal weight patients, weighing 60–79 kg has only 5 patients, with HTE 0.5 (SD = 0.326). Due to the limited number of patients in the normal weight category, it is not possible to draw any definitive conclusions.

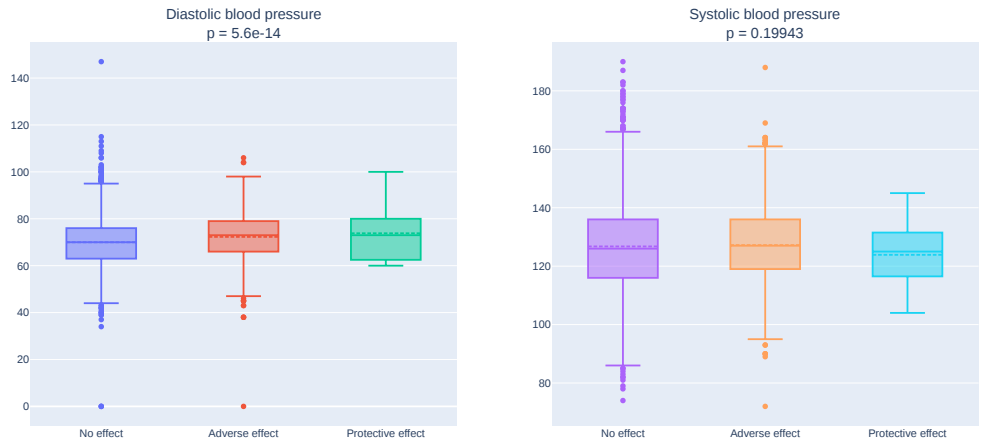

Figure 40: Distribution of HTEs according to blood pressure (diastolic, systolic in Vancomycin

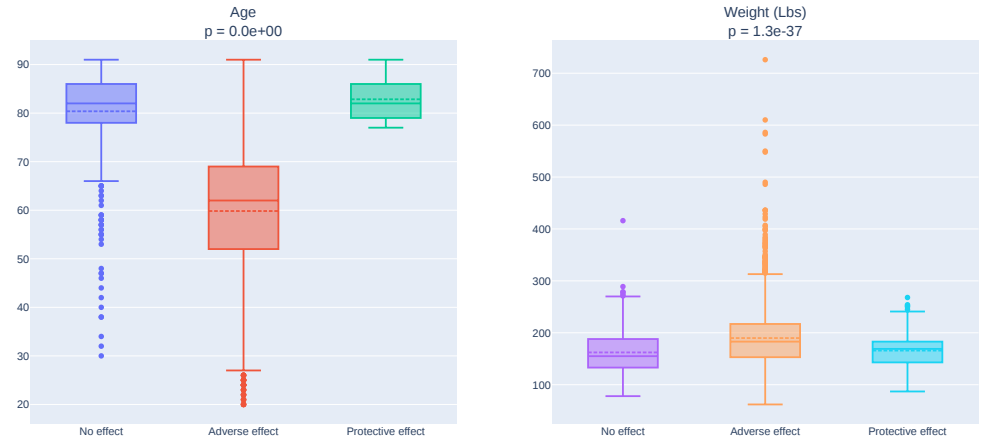

Figure 41: Distribution of HTEs according to age and weight in Lisinopril

Table 23 summarizes the mean HTE in the demographic categories. The highest mean HTE values, concerning an adequate number of patients, are detected to middle-aged and aged (40-79), females and overweight patients. The standard deviations are normal, so the evidence that was produced can be trusted.

Laboratory tests In creatinine and glucose, the HTEs follow the same pattern, AKI adverse effect presents significantly increased levels compared to no and protective effect. In bicarbonate, bun, sodium, and chloride, AKI adverse effect patients have significantly lower levels than the other two categories, $p < 0.01$, $p = 0.03$, $p < 0.01$ and $p = 0.0005$, respectively. Finally, in anion gap

Demographic	Category	N	HTE μ (σ)
Gender	Female	1615	0.5 ($\pm$ 0.4)
	Male	2152	0.45 ($\pm$ 0.35)
Age	18–39	192	0.6 ($\pm$ 0.2)
	40–59	1030	0.6 ($\pm$ 0.2)
	60–79	1857	0.55 ($\pm$ 0.3)
	$\geq$80	688	-0.07 ($\pm$ 0.2)
Weight	60–79 kg	5	0.5 ($\pm$ 0.326)
	80–99 kg	45	0.3 ($\pm$ 0.366)
	$\geq$100 kg	3717	0.46 ($\pm$ 0.35)

Table 23: Heterogeneous Treatment Effect (HTE) by Demographic Category in Lisinopril

protective effect patients present the highest level, adverse effect moderate and no effect the smaller ones (p = 0.035).

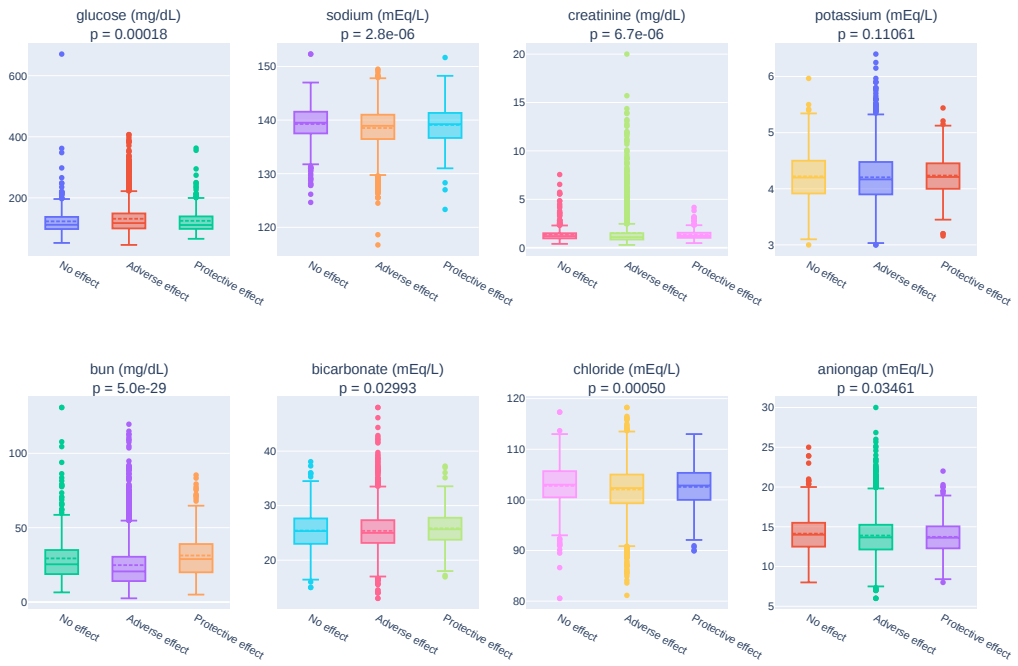

Figure 42: Distribution of HTEs according to different laboratory tests in Lisinopril

Glucose For lisinopril, HTE remains consistently high and positive across all glucose categories. In patients with normal fasting glucose ($\leq$ 100 mg/dL), the HTE mean is 0.44 (SD = 0.362; N = 951). Among those with impaired fasting glucose (100–125 mg/dL), the HTE mean is the same with the normal glucose values group at 0.44 (SD = 0.365; N = 1333). In patients with diabetes ($\geq$ 126 mg/dL), the HTE mean is the highest at 0.5 (SD = 0.334; N = 1483).

Laboratory Test	Category	N	HTE μ (σ)
Glucose	<100 mg/dL	951	0.44 ($\pm$ 0.36)
	100–125 mg/dL	1333	0.44 ($\pm$ 0.365)
	$\geq$ 126 mg/dL	1483	0.5 ($\pm$ 0.334)
Sodium	<135 mEq/L	537	0.463 ($\pm$ .35)
	135–145 mEq/L	3128	0.463 ($\pm$ 0.35)
	> 145 mEq/L	102	0.33 ($\pm$ 0.4)
Creatinine	<1.2 mg/dL	2081	0.5 ($\pm$ 0.343)
	1.2–1.9 mg/dL	1205	0.4 ($\pm$ 0.4)
	$\geq$2.0 mg/dL	481	0.5 ($\pm$ 0.3)
Potassium	<3.5 mEq/L	137	0.5 ($\pm$ 0.33)
	3.5-5.0 mEq/L	3433	0.46 ($\pm$ 0.35)
	>5.0 mEq/L	197	0.47 ($\pm$ 0.33)
BUN	<7 mg/dL	87	0.6 ($\pm$ 0.2)
	7–20 mg/dL	1538	0.6 ($\pm$ 0.3)
	¿20 mg/dL	2142	0.4 ($\pm$ 0.4)
Bicarbonate	<22 mEq/L	497	0.47 ($\pm$ 0.33)
	22–29 mEq/L	2719	0.46 ($\pm$ 0.35)
	¿29 mEq/L	551	0.44 ($\pm$ 0.36)
Chloride	<98 mEq/L	456	0.32 ($\pm$ 0.15)
	98–106	3127	0.23 ($\pm$ 0.16)
	>106	513	0.24 ($\pm$ 0.16)
Anion Gap	< 8 mEq/L	12	0.64 ($\pm$ 0.2)
	8–16 mEq/L	3042	0.45 ($\pm$ 0.35)
	>16 mEq/L	713	0.47 ($\pm$ 0.34)

Table 24: Lisinopril: Heterogeneous Treatment Effect (HTE) by Laboratory Category

Sodium Among hyponatremic ($<$ 135 mEq/L) and normonatremic (135–145 mEq/L) patients, HTE mean is the same, 0.463 (SD = 0.35), while the number of patients differs, with 537 and 3128, respectively. The hypernatremic category of patients presents the lowest HTE mean of 0.33 (SD = 0.38; N = 102).

Creatinine Most of the patients are classified in the normal creatinine group ($<$ 1.2 mg/dL), exhibited a highest HTE of 0.5 (SD = 0.343; N = 2081), whereas in the mildly elevated (1.2–1.9 mg/dL) group has a moderate mean HTE of 0.4 (SD=0.373; N = 1205). At severely elevated ($\geq$ 2.0 mg/dL) strata, which includes the lowest number of patients, HTE mean exhibits a high mean HTE of 0.48 (SD = 0.323; N=481).

Potassium In hyperkalemic ($<$ 3.5) and normokalemic (3.5–5.0) strata, HTEs were 0.47 (SD = 0.33; N = 197) and 0.46 (SD = 0.35; N = 3433), respectively. In hypokalemic patients, the mean HTE reaches the highest with 0.5 (SD = 0.33; N = 137)

BUN HTE demonstrates a clear positive trend as BUN levels increase. In patients with low BUN (below 7 mg/dL), the HTE mean is 0.6 (SD = 0.2; N = 67) as well as in patients with normal BUN levels of 7–20 mg/dL (SD = 0.3; N = 1538). The lowest HTE mean is observed in patients with elevated BUN ($>$ 20 mg/dL), where the value reaches 0.4 (SD = 0.4; N = 2142).

Bicarbonate Patients with metabolic acidosis ($<$ 22 mEq/L ; N = 1774) reached the highest mean HTE 0.47 (SD = 0.33; N = 497). Normal bicarbonate value (7-20 mEq/L) patients had almost the same mean HTE OD 0.46 (SD = 0.35; N = 2719). Finally, patients with metabolic alkalosis have the lowest HTE mean of 0.44 (SD = 0.36; N = 551).

Chloride Hypochloremics ($<$ 98 mEq/L) present the highest mean HTE of 0.32 (SD = 0.15; N = 456). The majority of patients belong to the normal chloride values (98-106 mEq/L) where the mean HTE is 0.23 (SD = 0.16; N = 3127). Hyperchloremics ($>$ 106 mEq/L) reach a mean HTE of 0.24 (SD = 0.16; N = 513).

Anion Gap Low values of anion gap ($<$ 8) detected to only 12 patients with mean HTE = 0.64 (SD = 0.2). In normal and elevated anion gap values the mean HTE is lower with 0.45 (SD = 0.35; N = 3042) and 0.47 (SD = 0.34; N = 713), respectively.

According to laboratory values presented in table 24, patients with high mean HTE are characterized by elevated levels of glucose ($\geq$ 126 mg/dL), and low level of potassium ($<$ 3.5 mEq/L), bicarbonate ($<$ 22 mEq/L), and chloride ($<$ 98 mEq/L).

Blood pressure In figure 43, only diastolic blood pressure presents statistical significance between the increased levels of AKI adverse effect patients compared to protective and no effect patients.

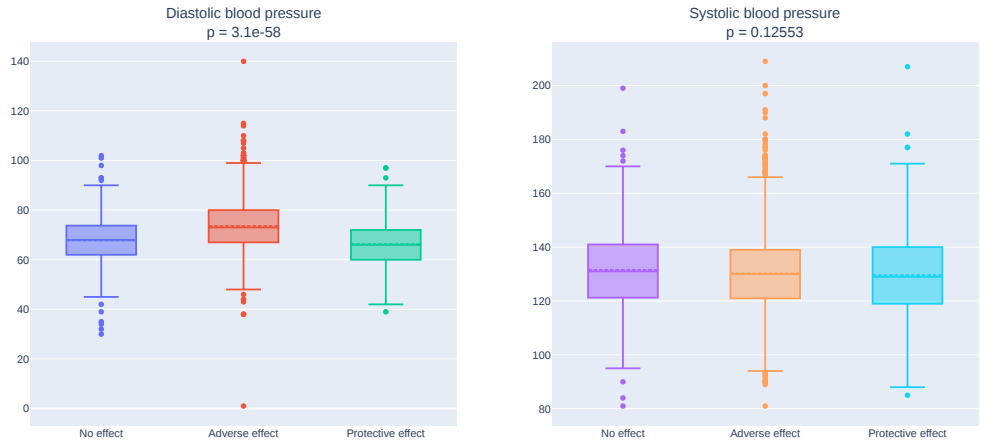

Figure 43: Distribution of HTEs according to blood pressure (diastolic, systolic in Lisinopril

The majority of the patients according to systolic blood pressure, belong to the normal category ($<$ 120) with mean HTE 0.46 (SD 0.353; N=3766) and only one patient with hypertension ($\geq$ 140) and a high HTE of 0.85. In diastolic blood pressure, there are no patients with normal values. In the first stage of hypertention (80–89) there are only 7 patients with a mean HTE 0.1 (SD = 0.35), and at the second stage of hypertension ($\geq$ 90), the mean HTE is 0.46 (SD 0.353; N=3760).

K.0.5. FUROSEMIDE

Demographics Furosemide-induced AKI patients are significantly ($p < 0.01$) older than patients with no AKI adverse effect and patients with negative or protective effect (Figure 44).

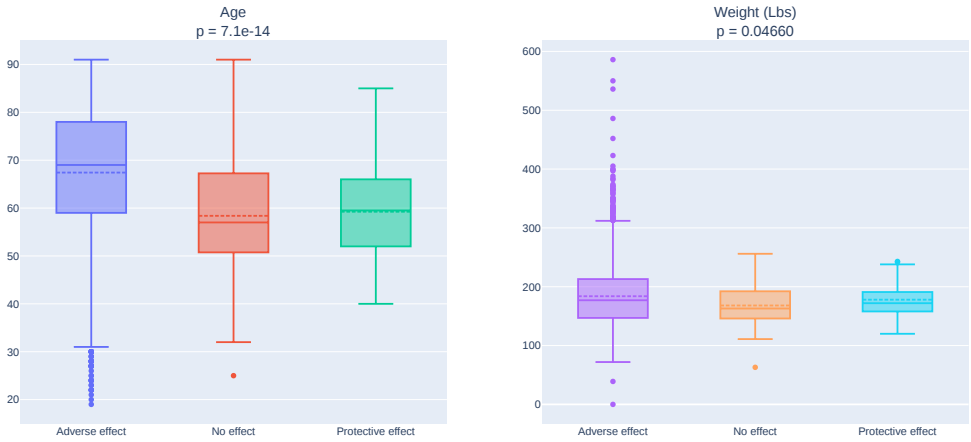

Figure 44: Distribution of HTEs according to age and weight in Furosemide

Age HTE mean is relatively high and stable across age groups, peaking at 0.59 in ≥80.

Demographic	Category	N	HTE μ (σ)
Gender	Female	1451	0.5 ($\pm$ 0.2)
	Male	1896	0.5 ($\pm$ 0.2)
Age	18–39	118	0.5 ($\pm$ 0.2)
	40–59	733	0.44 ($\pm$ 0.2)
	60–79	1739	0.5 ($\pm$ 0.2)
	≥80	757	0.6 ($\pm$ 0.2)
Weight	<60 kg	2	0.5 ($\pm$ 0.1)
	80–99 kg	46	0.6 ($\pm$ 0.15)
	≥100 kg	3293	0.5 ($\pm$ 0.2)

Table 25: Heterogeneous Treatment Effect (HTE) by Demographic Category in Furosemide

Gender Females (0.525) have a slightly higher HTE mean than males (0.495), with similar SDs, reflecting similar efficacy in both sexes.

Weight HTE mean is high across all weight categories, with the highest value observed in the 80–99 kg group (0.583), followed by <60 kg (0.543), and slightly lower in the ≥100 kg group (0.507). However, the sample size is extremely small in the ¡60 kg (N=2) and 80–99 kg (N=46) groups compared to the ≥100 kg group (N=3293), making the latter's estimate much more robust. The consistently high HTE mean across all weights suggests that furosemide's effect is substantial regardless of weight, but the limited data in lower weight groups preclude any strong conclusions about increased effect per body mass. The most reliable estimate is for the ≥100 kg group, where the effect remains high.

Table 25 summarizes the mean HTE in the demographic categories. The highest mean HTE values, concerning an adequate number of patients, are detected in the aged (60-79) and over 80,

males and females and overweight patients. The standard deviations are small, so the evidence produced is robust and can be trusted.

Laboratory tests In the figure 45, it is interesting to highlight the significantly (p = 0.002) higher level of sodium levels in patients with AKI adverse effect, compare to the other two categories. In bun, patients with AKI adverse effects present significantly (p=0.001) high levels than no and protective effect patients. Furthermore, furosemide-induced AKI patients present a significant (p = 0.00004) increased level of bicarbonate.

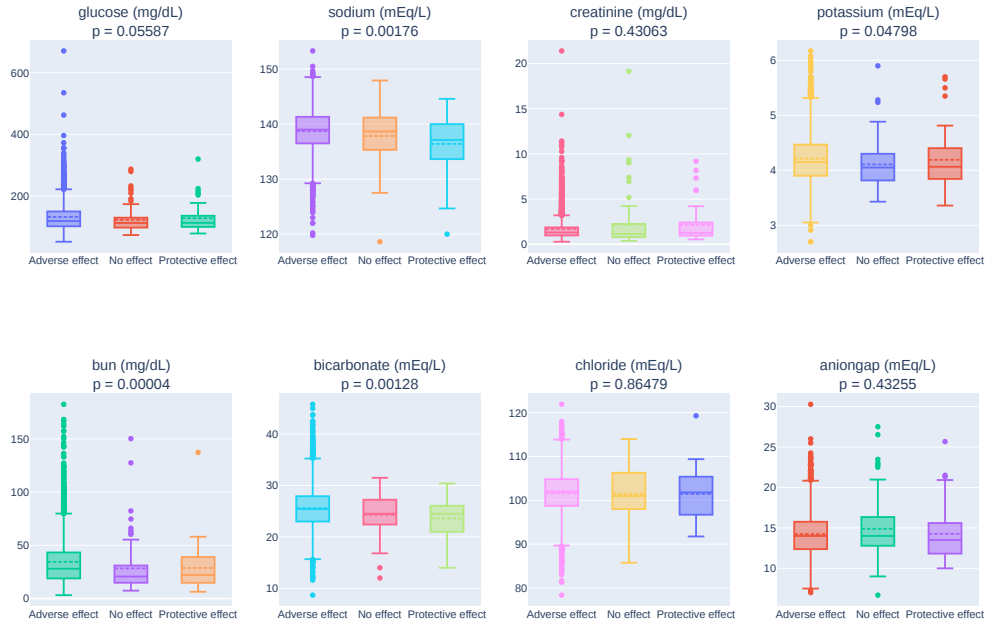

Figure 45: Distribution of HTEs according to different laboratory tests in Furosemide

Glucose HTE mean remains high and remarkably stable across all categories of glucose metabolism. In patients with normal fasting glucose (<100 mg/dL), the HTE mean is 0.516 (SD = 0.216; N=701). Among those with impaired fasting glucose (100–125 mg/dL, pre-diabetes), the HTE mean is 0.510 (SD = 0.209; N=1233). In patients with diabetes (glucose $\geq$ 126 mg/dL), the HTE mean is 0.501 (SD 0.187; N=1413). Notably, the standard deviation decreases slightly with increasing glucose, suggesting that the effect of furosemide becomes somewhat more consistent among individuals with higher glucose levels.

Sodium In patients with hyponatremia (serum sodium <135 mEq/L), the HTE mean is 0.485 (SD = 0.2; N=514), while it is 0.511 ((SD = 0.2; N=2704) in individuals with normal sodium levels (135–145 mEq/L), and 0.519 (SD = 0.2; N=129) in those experiencing hypernatremia (>145 mEq/L). This data demonstrates that furosemide's impact maintains a high level of effectiveness across all sodium concentrations.

Creatinine Furosemide shows consistently high HTE across all creatinine levels, with slightly increased efficacy in patients with mildly elevated creatinine (1.2–1.9 mg/dL) with mean HTE of

Laboratory Test	Category	N	HTE μ (σ)
Glucose	<100 mg/dL	701	0.52 ($\pm$ 0.22)
	100–125 mg/dL	1233	0.51 ($\pm$ 0.21)
	$\geq$126 mg/dL	1413	0.50 ($\pm$ 0.19)
Sodium	<135 mEq/L	514	0.49 ($\pm$ 0.2)
	135–145 mEq/L	2704	0.51 ($\pm$ 0.2)
	>145 mEq/L	129	0.52 ($\pm$ 0.2)
Creatinine	<1.2 mg/dL	1407	0.50 ($\pm$ 0.21)
	1.2–1.9 mg/dL	1183	0.53 ($\pm$ 0.19)
	$\geq$2.0 mg/dL	757	0.49 ($\pm$ 0.21)
Potassium	<3.5 mEq/L	90	0.513 ($\pm$ 0.197)
	3.5–5.0 mEq/L	3042	0.507 ($\pm$ 0.201)
	>5.0 mEq/L	215	0.511 ($\pm$ 0.217)
BUN	<7 mg/dL	19	0.380 ($\pm$ 0.256)
	7–20 mg/dL	962	0.482 ($\pm$ 0.221)
	>20 mg/dL	2366	0.519 ($\pm$ 0.192)
Bicarbonate	<22 mEq/L	555	0.487 ($\pm$ 0.2)
	22–29 mEq/L	2190	0.508 ($\pm$ 0.2)
	>29 mEq/L	602	0.525 ($\pm$ 0.2)
Chloride	<98 mEq/L	691	0.494 ($\pm$ 0.2)
	98–106 mEq/L	2055	0.516 ($\pm$ 0.2)
	>106 mEq/L	601	0.494 ($\pm$ 0.2)
Anion Gap	< 8 mEq/L	9	0.34 ($\pm$ 0.2)
	8–16 mEq/L	2546	0.51 ($\pm$ 0.2)
	>16 mEq/L	792	0.503 ($\pm$ 0.2)

Table 26: Furosemide: Heterogeneous Treatment Effect (HTE) by Laboratory Category

0.526 (SD = 0.19; N=1183) compared to those with normal ($<$ 1.2 mg/dL) where HTE mean is 0.5 (SD = 0.21; N=1407) or severely elevated levels ($\geq$ 2.0 mg/dL) where the mean HTE is 0.493 (SD = 0.21; N=757).

Potassium Furosemide exhibits notably high HTE across all potassium categories, with the strongest effect observed in hypokalemic patients ($<$ 3.5 mEq/L) where mean HTE is 0.513 (SD = 0.197; N=90), and a slightly lower but still substantial effect in normal values of potassium (3.5–5.0 mEq/L) the HTE is 0.507 (SD = 0.201; N=3042) and hyperkalemia patients ($>$ 5.0 mEq/L) have HTE mean 0.511 (SD = 0.217; N=215).

BUN BUN levels rise, furosemide displays an increasingly stronger mean HTE. For patients with low BUN ($<$ 7 mg/dL), the HTE registers at a moderate 0.380 and exhibits broader variability (SD = 0.256; N = 19). Patients with normal BUN levels (7–20 mg/dL) experience an enhanced HTE of 0.482 (SD = 0.221; N = 962), while those with elevated BUN ($>$ 20 mg/dL) reach an HTE of 0.519, alongside decreased variability (SD = 0.192; N = 2366).

Bicarbonate HTE of bicarbonate levels, showing a notable difference between metabolic acidosis and alkalosis states. In patients with low bicarbonate ($<$ 22 mEq/L), indicative of metabolic acidosis, the HTE mean is moderately high at 0.487 (SD = 0.224; N = 555). Conversely, in patients with elevated bicarbonate ($>$ 29 mEq/L), representing metabolic alkalosis, furosemide demonstrates

a higher and more consistent effect of mean HTE 0.525 (SD = 0.172; N = 602). Finally, patients with normal bicarbonate values present a mean HTE of 0.508 (SD = O.2; N = 2190).

Chloride Furosemide exhibits a consistent mean HTE across varying chloride levels. In hypochloremic patients (< 98 mEq/L), the HTE mean is 0.494 (SD = 0.207; N = 691), while in hyperchloremic patients (> 106 mEq/L), it remains nearly identical at 0.494 (SD = 0.213; N = 601). In normal chloride levels (98–106 mEq/L), the HTE mean reaches the highest value of 0.516 (SD = 0.2;N = 2055).

Anion gap HTE mean is fairly consistent across anion gap categories. Patients with normal anion gap (8–16 mEq/L) show an HTE mean of 0.51 (SD = 0.2; N = 2546), while those with elevated anion gap (¿16 mEq/L) have a slightly lower HTE of 0.50 (SD = 0.2; N = 792). A small subgroup with low anion gap (¡8 mEq/L) exhibits a lower HTE of 0.34 (SD = 0.2; N = 9).

According to laboratory values presented in table 26, patients with high mean HTE are characterized from elevated levels of BUN (> 20 mg/dL), bicarbonate (> 29 mEq/L), and sodium (> 145 mEq/L), low levels of glucose (< 100 mg/dL), potassium (< 3.5 mEq/L) and normal levels of creatinine (1.2–1.9 mg/dL), chloride (98–106 mEq/L) and anion gap (8–16 mEq/L).

Blood pressure Furosemide-induced AKI patients have lower diastolic and higher systolic blood pressure than patients with no DAKI effect or with protective effect (Figure 46).

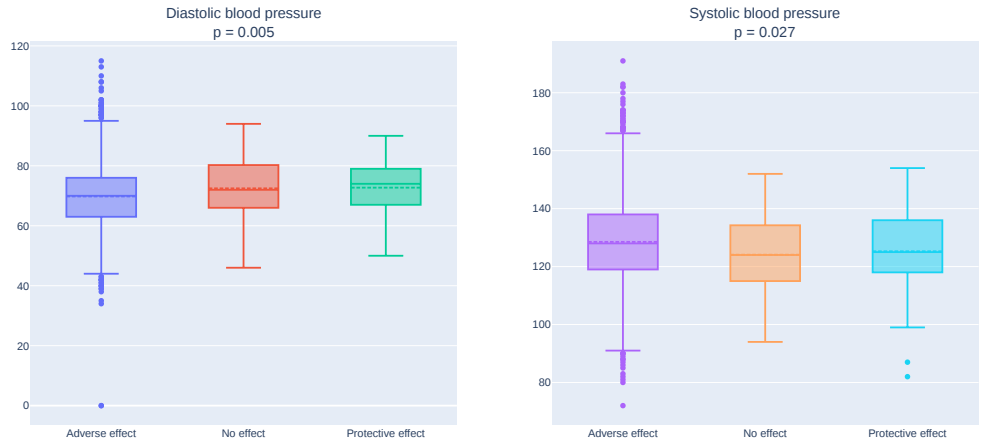

Figure 46: Distribution of HTEs according to blood pressure (diastolic, systolic in Furosemide

Furosemide exhibits a strong and consistent antihypertensive effect across systolic and diastolic blood pressure categories. In normotensive patients (systolic <120 mmHg), the HTE mean is 0.508 (SD = 0.202; N = 3347), indicating a substantial average reduction in systolic blood pressure. For diastolic blood pressure, the data show a very high HTE mean of 0.578 in patients with diastolic <80 mmHg, but this is based on a single patient and cannot be interpreted reliably. In stage 1 hypertension (80–89 mmHg), the HTE mean is 0.371 (SD = 0.293, N = 12), and in stage 2 hypertension (≥90 mmHg), it is 0.508 (SD = 0.201; N = 3334), mirroring the strong effect seen in systolic blood pressure.

K.0.6. PANTOPRAZOLE

Demographics The age of patients with pantoprazole-induced AKI is significant (p = 0.04) lower than patients with no DAKI effect and with a protective effect (Figure 47).

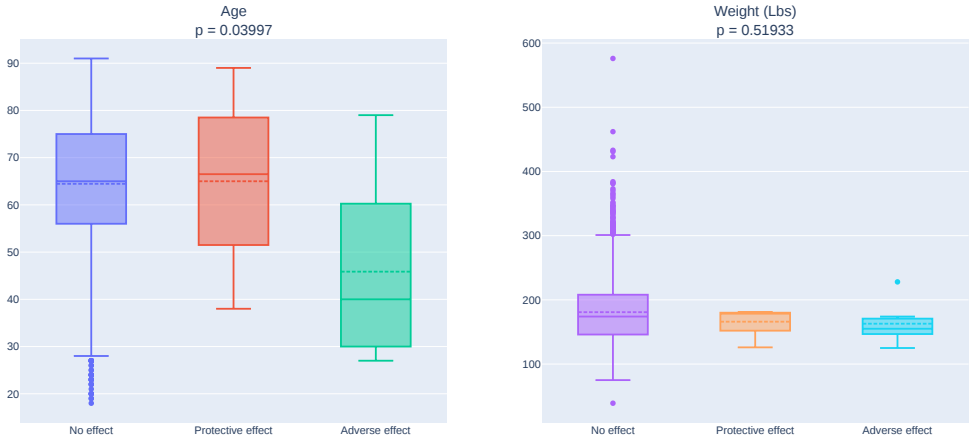

Figure 47: Distribution of HTEs according to age and weight in Pantoprazole

Age HTE mean is close to zero in all age groups, suggesting little average effect or possibly a neutral risk/benefit profile for the measured outcome. SDs are low, indicating consistency.

Demographic	Category	N	HTE μ (σ)
Gender	Female	1104	0.002 ($\pm$ 0.1)
	Male	1581	0.001 ($\pm$ 0.06)
Age	18–39	138	0.01 ($\pm$ 0.2)
	40–59	743	0 ($\pm$ 0.05)
	60–79	1325	0 ($\pm$ 0.05)
	$\geq$80	479	0 ($\pm$ 0.06)
Weight	80–99 kg	45	0 ($\pm$ 0)
	$\geq$100 kg	2638	0.001 ($\pm$ 0.065)

Table 27: Heterogeneous Treatment Effect (HTE) by Demographic Category in Pantoprazole

Gender Both sexes have near-zero HTE means, supporting the age findings. Females showed a slightly greater CATE than males.

Weight HTE mean is essentially zero across all weight categories, with the only substantial sample size in the $\geq$100 kg group with mean HTE 0.001 (SD = 0.065; N = 2638). The lower weight categories ($<$60 kg, 60–79 kg, 80–99 kg) have extremely small sample sizes (N=1 or N=45) and HTE means of zero, offering no reliable information about treatment effect in these groups. The data indicate that pantoprazole has no meaningful heterogeneous treatment effect across weight categories, and the only robust estimate ($\geq$100 kg) confirms a negligible effect.

Table 27 presents the mean HTE in the demographic categories. The mean HTE values for every demographic category are close to 0 or 0 which reflects the lack of pantoprazole-induced AKI effect in our data.

Laboratory tests As it is obvious from figure 48, only bicarbonate presents significant results, as pantoprazole-induced AKI patients seem to have lower levels of bicarbonate compared to patients with no AKI adverse effect and protective effect.

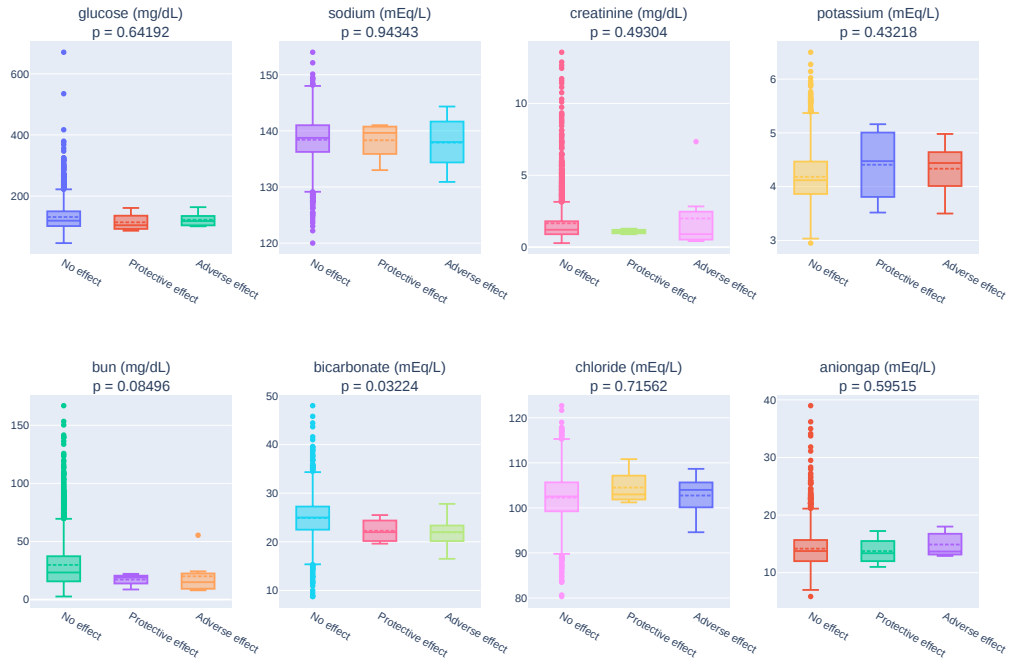

Figure 48: Distribution of HTEs according to different laboratory tests in Pantoprazole

Across all laboratory categories—including glucose, sodium, creatinine, potassium, BUN, bicarbonate, chloride, and anion gap—the HTE means for pantoprazole are essentially zero, with only minimal variation. For example, in patients with normal fasting glucose, the HTE mean is 0.004 (SD 0.080, N=931), and in those with diabetes, it is 0.001 (SD 0.051, N=1156). Similarly, for sodium, the HTE mean is 0.002 (SD 0.084, N=425) in hyponatremia and 0.001 (SD 0.061, N=2179) in normonatremia. These consistently negligible values indicate that pantoprazole does not exert a clinically meaningful effect on these laboratory parameters or on the primary outcome measured in the study population (Table 28).

Blood Pressure Figure 49 do not show any significance in different values of blood pressure between the three group of patients.

Pantoprazole's HTE mean is 0.001 (SD 0.064, N=2685) in normal blood pressure patients, indicating a neutral effect, consistent with its known safety profile and lack of cardiovascular impact. For diastolic ¡80 mmHg, the HTE mean is 0 (N=2), and in stage 2 hypertension, it is 0.001 (SD 0.064, N=2680), again showing no significant effect.

Laboratory Test	Category	N	μ (σ)
Glucose	<100 mg/dL	598	-0.003 ($\pm$ 0.1)
	100–125 mg/dL	931	0.004 ($\pm$0.1)
	$\geq$126 mg/dL	1156	0.001 ($\pm$ 0.05)
Sodium	<135 mEq/L	425	0.002 ($\pm$ 0.1)
	135–145 mEq/L	2179	0.001 ($\pm$ 0.1)
	>145 mEq/L	81	0 ($\pm$ 0)
Creatinine	<1.2 mg/dL	1278	0.001 ($\pm$ 0.1)
	1.2–1.9 mg/dL	835	0 ($\pm$ 0.05)
	$\geq$2.0 mg/dL	572	0.003 ($\pm$ 0.1)
Potassium	<3.5 mEq/L	109	0 ($\pm$ 0)
	3.5–5.0 mEq/L	2428	0.002 ($\pm$ 0.1)
	>5.0 mEq/L	148	-0.007 ($\pm$ 0.1)
BUN	<7 mg/dL	53	0 ($\pm$ 0)
	7–20 mg/dL	1005	0.002 ($\pm$ 0.1)
	>20 mg/dL	1627	0.001 ($\pm$ 0.04)
Bicarbonate	<22 mEq/L	545	0.002 ($\pm$ 0.1)
	22–29 mEq/L	1777	0.001 ($\pm$ 0.1)
	>29 mEq/L	363	0 ($\pm$ 0)
Chloride	<98 mEq/L	478	0.002 ($\pm$ 0.05)
	98–106 mEq/L	1580	0.001 ($\pm$ 0.1)
	>106 mEq/L	627	0.002 ($\pm$ 0.1)
Anion Gap	8–16 mEq/L	2070	0.0005 ($\pm$ 0.1)
	<8 mEq/L	6	0 ($\pm$ 0)
	>16 mEq/L	609	0.003 ($\pm$ 0.1)

Table 28: Heterogeneous Treatment Effect (HTE) estimates for Pantoprazole across various laboratory value categories.

K.0.7. OMEPRAZOLE

Demographics Omeparzole-induced AKI patients belong to a significant ($p < 0.01$) younger age group than patients with no DAKI effect and patients with protective effect (Figure 50).

Age HTE means are very close to zero across all ages, with low SDs, indicating little individual variation in effect.

Gender Females have HTE mean equal to 0.007 (SD 0.112, N=3229), while for males it is -0.003 (SD 0.085, N=3011). This indicates that the estimated heterogeneous treatment effect is very close to zero for both genders, with a slight positive value in females and a slight negative value in males. The difference between the groups is minimal (0.01), and both standard deviations suggest moderate variability within each gender group. Thus, there is no clinically meaningful difference in the heterogeneous treatment effect of omeprazole between males and females, and the effect is essentially neutral in both populations.

Weight HTE of omeprazole is consistently close to zero across all weight categories. In the lowest weight groups (<60 kg and 60–79 kg), the HTE mean is exactly zero, but these groups have

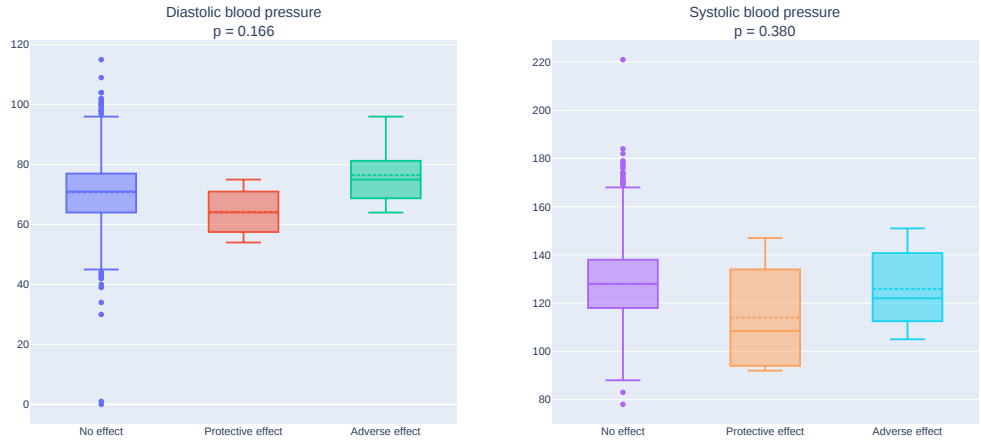

Figure 49: Distribution of HTEs according to blood pressure (diastolic, systolic) in Pantoprazole

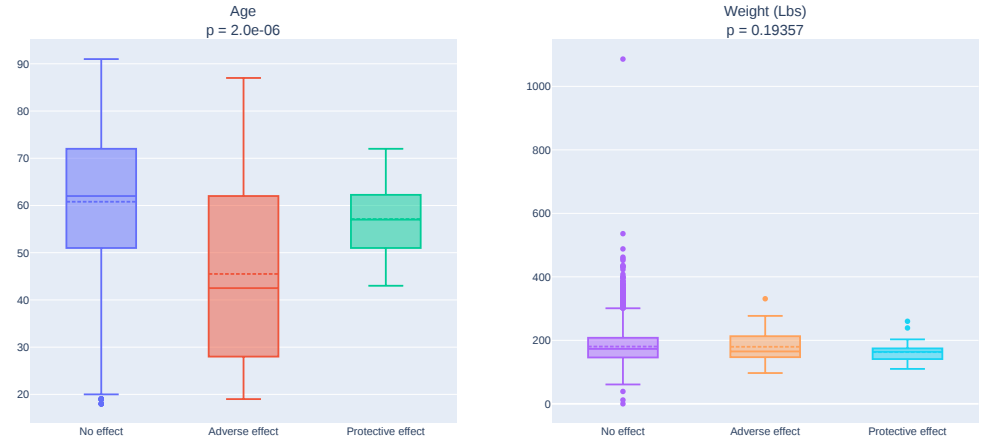

Figure 50: Distribution of HTEs according to age and weight in Omeprazole

extremely small sample sizes, limiting interpretability. In the 80–99 kg and $\geq$100 kg groups, which have larger sample sizes (especially $\geq$100 kg), the HTE means remain very close to zero (0.011 and 0.002, respectively), with similar standard deviations ($\tilde{0}$.10), indicating low variability and a stable, negligible effect regardless of body weight. This pattern suggests that omeprazole's effectiveness or risk profile does not meaningfully differ by weight, and the drug's impact is essentially neutral across the weight spectrum.

Demographic	Category	N	HTE μ (σ)
Gender	Female	3229	0.007 ($\pm$ 0.1)
	Male	3011	-0.003 ($\pm$ 0.1)
Age	18–39	655	0.02 ($\pm$ 0.15)
	40–59	1943	0 ($\pm$ 0.1)
	60–79	2787	0 ($\pm$ 0.1)
	$\geq$80	855	0 ($\pm$ 0.05)
Weight	$<$60 kg	3	0 ($\pm$ 0)
	60–79 kg	9	0 ($\pm$ 0)
	80–99 kg	91	0.01 ($\pm$ 0.1)
	$\geq$100 kg	6137	0.002 $\pm$ 0.1

Table 29: Heterogeneous Treatment Effect (HTE) by Demographic Category in Omeprazole

Table 29 presents the mean HTE in the demographic categories. The mean HTE values for every demographic category are close to 0 or 0 which reflects the lack of omeprazole-induced AKI effect in our data.

Laboratory tests In laboratory tests, glucose, creatinine, bun and bicarbonate have similar patterns as AKI adverse effect patients have significant lower levels in all these tests than no DAKI effect patient and patients with omeprazole protective effect, with p = 0.0001, p = 0.001, p $<$ 0.01 and p = 0.0001, respectively. Moreover, in chloride levels patients with omeprazole-induced AKI are significant (0.003) higher than the other two groups (Figure 51).

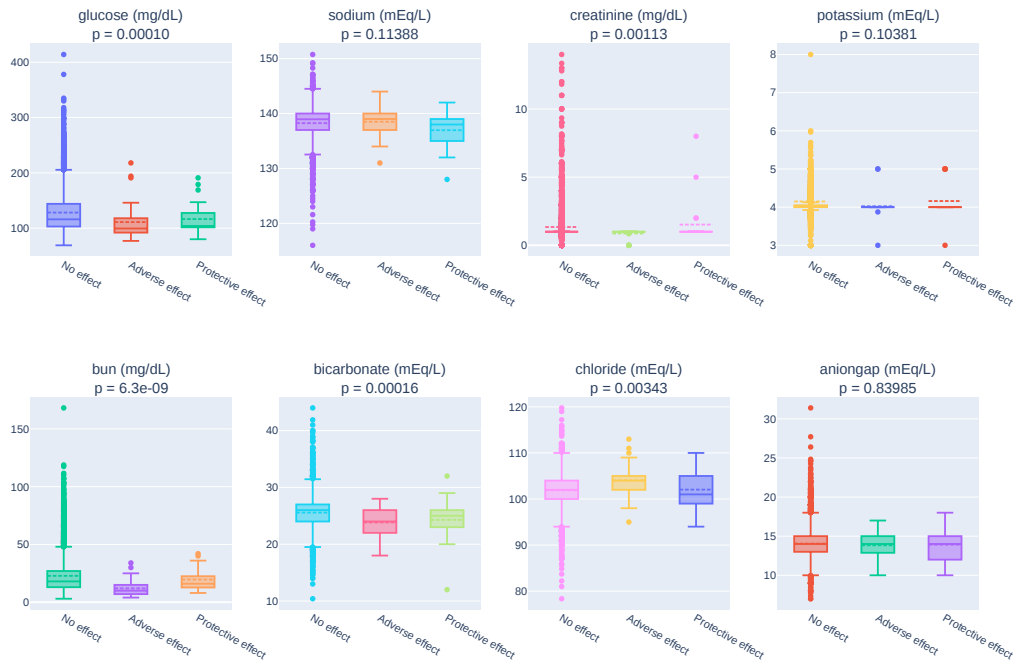

Figure 51: Distribution of HTEs according to different laboratory tests in Omeprazole

HTE means are consistently close to zero across all laboratory categories, indicating a neutral effect profile. For example, in patients with normal fasting glucose, the HTE mean is -0.003 (SD = 0.058; N = 598), and in those with normonatremia, the HTE mean is 0.003 (SD = 0.101; N = 5535). In patients with a normal anion gap, the HTE mean is 0.002 (SD = 0.1; N = 5065), and even among those with laboratory abnormalities such as hyponatremia where HTE mean is -0.003 (SD = 0.1; N = 664) or metabolic acidosis where HTE mean is 0.008 (SD = 0.145, N = 475), the values remain very small. Considering blood urea nitrogen, the HTE mean is 0.091 (SD = 0.289; N=88) in patients with low BUN, 0.003 (SD = 0.109; N = 3433) in those with normal BUN, and -0.001 (SD = 0.072, N = 2719) in those with elevated BUN. These results further support the conclusion that omeprazole does not have a clinically significant impact on metabolic or electrolyte parameters, regardless of renal function. This neutral effect is consistent with the established safety profile of omeprazole in large, diverse populations (Table 30).

Laboratory Test	Category	N	HTE μ (σ)
Glucose	$<$100 mg/dL	1166	0.012 ($\pm$ 0.1)
	100–125 mg/dL	2701	-0.001 ($\pm$ 0.1)
	$\geq$126 mg/dL	2373	0.001 ($\pm$ 0.1)
Sodium	$<$135 mEq/L	664	-0.003 ($\pm$ 0.1)
	135–145 mEq/L	5535	0.003 ($\pm$ 0.1)
	$>$145 mEq/L	41	0 ($\pm$ 0)
Creatinine	$<$1.2 mg/dL	4887	0.003 ($\pm$ 0.1)
	1.2–1.9 mg/dL	380	0 ($\pm$ 0)
	$\geq$2.0 mg/dL	973	-0.004 ($\pm$ 0.1)
Potassium	$<$3.5 mEq/L	75	0 ($\pm$ 0.2)
	3.5–5.0 mEq/L	5388	0.003 ($\pm$ 0.1)
	$>$5.0 mEq/L	777	-0.004 ($\pm$ 0.1)
BUN	$<$7 mg/dL	88	0.091 ($\pm$ 0.3)
	7–20 mg/dL	3433	0.003 ($\pm$ 0.1)
	$>$20 mg/dL	2719	-0.001 ($\pm$ 0.1)
Bicarbonate	$<$22 mEq/L	475	0.008 ($\pm$ 0.145)
	22–29 mEq/L	5065	0.002 ($\pm$ 0.1)
	$>$29 mEq/L	700	-0.003 ($\pm$ 0.05)
Chloride	$<$98 mEq/L	720	-0.003 ($\pm$ 0.1)
	98–106 mEq/L	4681	0.002 ($\pm$ 0.1)
	$>$106 mEq/L	839	0.005 ($\pm$ 0.1)
Anion Gap	8–16 mEq/L	4970	0.002 ($\pm$ 0.1)
	$<$8 mEq/L	5	0 ($\pm$ 0)
	$>$16 mEq/L	1265	0.001 ($\pm$ 0.1)

Table 30: Omeprazole: Heterogeneous Treatment Effect (HTE) by Laboratory Category

Blood Pressure Figure 52 shows a significant result only in systolic blood pressure, where AKI adverse effect patients have lower values than no DAKI effect patients and higher values than protective effect patients.

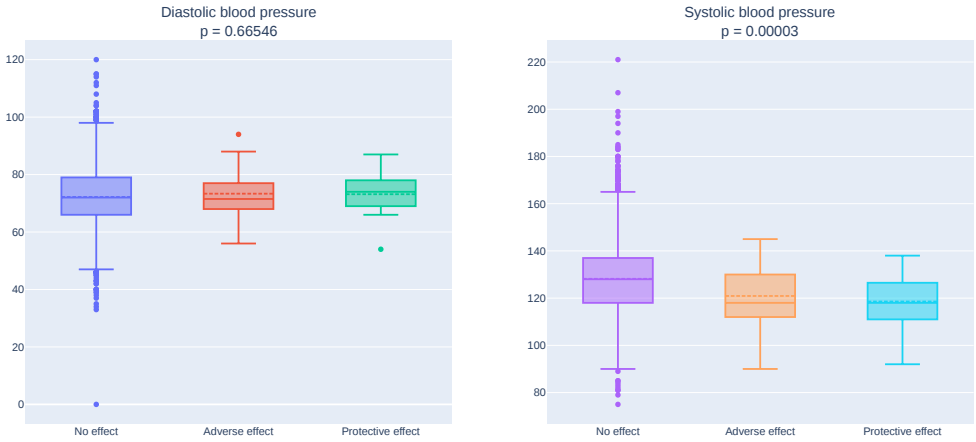

Figure 52: Distribution of HTEs according to blood pressure (diastolic, systolic) in Omeprazole

In systolic blood pressure, HTE mean is 0.002 (SD = 0.1; N = 6239), reflecting a neutral effect in normotensive patients. For diastolic ¡80 mmHg, the HTE mean is 0 (N = 1), and in stage 2 hypertension, it is 0.002 (SD = 0.101; N = 6230).

K.0.8. ALLOPURINOL

Demographics Figure 53 presents significant (p = 0.0001) results only in different weights between patients without DAKI effect and patients with protective effect, with the first ones to be heavier than the last ones.

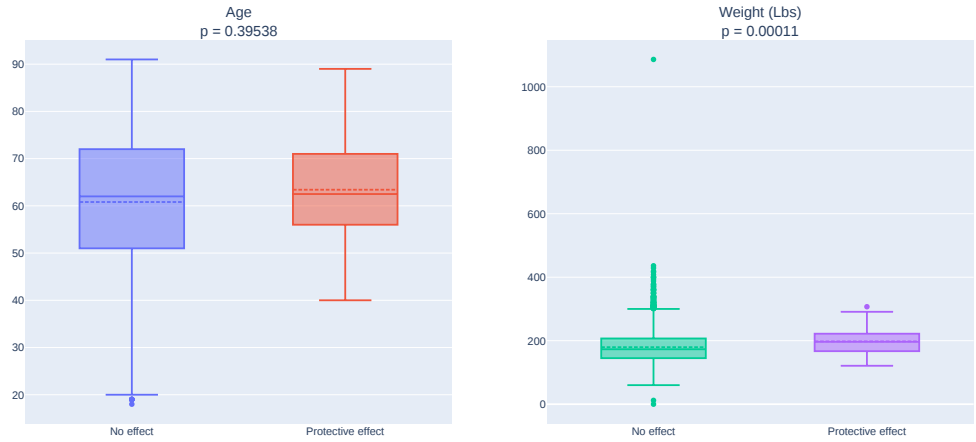

Figure 53: Distribution of HTEs according to age and weight in Allopurinol

Age HTE mean is zero or slightly negative across all age groups, suggesting little average effect or a possible slight reduction in the measured outcome. SDs are moderate, indicating some variability.

Demographic	Category	N	HTE μ (σ)
Gender	Female	1907	-0.008 ($\pm$ 0.1)
	Male	2161	-0.025 ($\pm$ 0.2)
Age	18–39	396	0 ($\pm$ 0)
	40–59	1241	-0.02 ($\pm$ 0.1)
	60–79	1927	-0.02 ($\pm$ 0.15)
	$\geq$80	504	-0.01 ($\pm$ 0.1)
Weight	$<$60 kg	2	0 ($\pm$ 0)
	60–79 kg	11	0 ($\pm$ 0)
	80–99 kg	47	0 ($\pm$ 0)
	$\geq$100 kg	4008	-0.017 ($\pm$ 0.1)

Table 31: Heterogeneous Treatment Effect (HTE) by Demographic Category in Allopurinol

Gender HTE mean for females is -0.008 (SD = 0.088; N = 1907), and for males it is -0.025 (SD = 0.158, N = 2161). Both values are negative and close to zero, indicating that, on average, allopurinol may have a very slight negative effect on the measured outcome in both genders. The difference between males and females is minimal (a difference of 0.017), and the standard deviations suggest moderate variability in individual responses, with somewhat greater variability among males.

Weight For allopurinol, the HTE means are exactly zero in the lower weight categories ($<$60 kg, 60–79 kg, 80–99 kg), but these groups have very small sample sizes (N=2, 11, and 47, respectively), making these estimates unreliable and not generalizable. In the $\geq$100 kg group—the only category with a large and robust sample size (N=4008), the HTE mean is slightly negative at -0.017 with a standard deviation of 0.131. This small negative value indicates a very modest average effect, which is not likely to be clinically significant given both the magnitude and the variability. Overall, these results suggest that the heterogeneous treatment effect of allopurinol does not vary meaningfully across weight categories.

Table 31 presents the mean HTE in the demographic categories. The mean HTE values for every demographic category are close to 0 (negative and positive), 0 which reflects the lack of allopurinol-induced AKI effect in our data.

Laboratory tests As it is obvious from figure 54, only bicarbonate and creatinine present significant results. In creatinine (p = 0.03), patients without DAKI effect have higher levels than patients with protective effect, but in bicarbonate (p = 0.03) is the opposite.

Allopurinol's HTE means are also near zero or slightly negative across all laboratory categories. For example, in patients with normal fasting glucose, the HTE mean is -0.008 (SD = 0.09; N = 744), while in those with diabetes, it is -0.021 (SD = 0.143; N = 1495). For sodium, the HTE mean is -0.005 (SD = 0.071; N = 399) in hyponatremia and -0.019 (SD = 0.136; N = 3632) in normonatremia. Across creatinine and BUN categories, the HTE means remain close to zero, with minor negative values that are not clinically significant. The slight negative HTE values observed in our data for glucose, sodium, or other metabolic variables are small and they are not create strong evidence about allopurinol's effect based on different levels of laboratory values (Table 32).

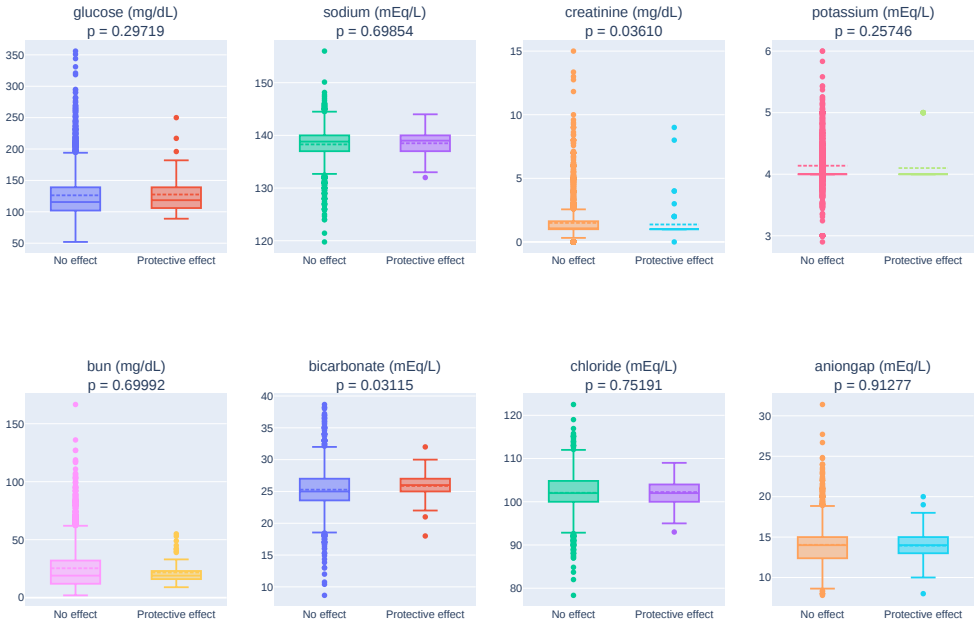

Figure 54: Distribution of HTEs according to different laboratory tests in Allopurinol

Blood Pressure Patients with a protective effect have significantly higher blood pressure (systolic, diastolic) than patients without the DAKI effect. In systolic blood pressure p-value is equal to 0.03 and in diastolic p-value is equal to 0.0004 (Figure 55).

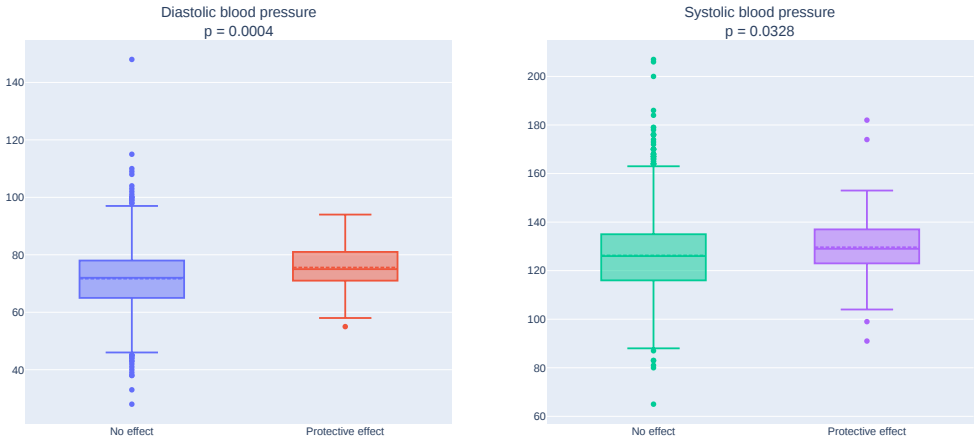

Figure 55: Distribution of HTEs according to blood pressure (diastolic, systolic) in Allopurinol

Laboratory Test	Category	N	HTE μ (σ)
Glucose	<100 mg/dL	744	-0.01 ($\pm$0.1)
	100–125 mg/dL	1829	-0.018 ($\pm$ 0.1)
	$\geq$126 mg/dL	1495	-0.021 ($\pm$ 0.1)
Sodium	<135 mEq/L	399	-0.005 ($\pm$ 0.1)
	135–145 mEq/L	3632	-0.019 ($\pm$ 0.1)
	>145 mEq/L	37	0 ($\pm$ 0)
Creatinine	<1.2 mg/dL	2896	-0.021 ($\pm$ 0.1)
	1.2–1.9 mg/dL	264	0 ($\pm$ 0)
	$\geq$2.0 mg/dL	908	-0.010 ($\pm$ 0.1)
Potassium	<3.5 mEq/L	63	0 ($\pm$ 0)
	3.5–5.0 mEq/L	3536	-0.018 ($\pm$ 0.1)
	>5.0 mEq/L	469	-0.015 ($\pm$ 0.1)
BUN	<7 mg/dL	79	0 ($\pm$ 0)
	7–20 mg/dL 2052	-0.019 ($\pm$0.1)	
	>20 mg/dL	1937	-0.016 ($\pm$ 0.1)
Bicarbonate	<22 mEq/L	377	-0.005 ($\pm$ 0.1)
	22–29 mEq/L	3237	-0.019 ($\pm$ 0.1)
	>29 mEq/L	454	-0.011 ($\pm$ 0.1)
Chloride	<98 mEq/L	454	-0.011 ($\pm$ 0.1)
	98–106 mEq/L	3003	-0.020 ($\pm$ 0.1)
	>106 mEq/L	611	-0.010 ($\pm$ 0.1)
Anion Gap	8–16 mEq/L	3202	-0.018 ($\pm$ 0.1)
	<8 mEq/L	1	0 ($\pm$ 0)
	>16 mEq/L	865	-0.013 ($\pm$ 0.1)

Table 32: Allopurinol: Heterogeneous Treatment Effect (HTE) by Laboratory Category

Allopurinol's HTE mean is -0.017 (SD 0.130, N=4067) in patients with normal blood pressure, indicating a very slight negative effect, which is not clinically significant. This is consistent with studies showing allopurinol has no major effect on blood pressure in most populations. For diastolic ¡80 mmHg, the HTE mean is 0 (N=1), and in stage 2 hypertension, it is -0.017 (SD 0.130, N=4056), again showing a neutral profile.