# Systematic Estimation of Treatment Effect on Hospitalization Risk as a Drug Repurposing Screening Method

Costa Georgantas[†1], Jaume Banus[1], Roger Hullin[2] and Jonas Richiardi[1]

[1]*Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland*

[2]*Department of Cardiology, Lausanne University Hospital, Lausanne, Switzerland*

[†]*E-mail: costa.georgantas@chuv.ch*

Drug repurposing (DR) intends to identify new uses for approved medications outside their original indication. Computational methods for finding DR candidates usually rely on prior biological and chemical information on a specific drug or target but rarely utilize real-world observations. In this work, we propose a simple and effective systematic screening approach to measure medication impact on hospitalization risk based on large-scale observational data. We use common classification systems to group drugs and diseases into broader functional categories and test for non-zero effects in each drug-disease category pair. Treatment effects on the hospitalization risk of an individual disease are obtained by combining widely used methods for causal inference and time-to-event modelling. 6468 drug-disease pairs were tested using data from the UK Biobank, focusing on cardiovascular, metabolic, and respiratory diseases. We determined key parameters to reduce the number of spurious correlations and identified 7 statistically significant associations of reduced hospitalization risk after correcting for multiple testing. Some of these associations were already reported in other studies, including new potential applications for cardioselective beta-blockers and thiazides. We also found evidence for proton pump inhibitor side effects and multiple possible associations for anti-diabetic drugs. Our work demonstrates the applicability of the present screening approach and the utility of real-world data for identifying potential DR candidates.

*Keywords*: Drug repurposing; Propensity score matching; Cox regression; Real-world data

## 1. Introduction

Drug discovery is a rarely successful and extremely costly process that can span decades before commercialization. Drug repurposing (DR), or re-utilizing an existing medication for another use, has the potential to cut down the cost of development by a factor of 10.[1] DR is still dependent on clinical trial success and only approximately 30% of repurposed drugs go from phase I to market,[2] a process that can take multiple years. The majority of trials fail due to insufficient efficacy or the existence of other superior alternatives. Computational methods can reduce the chances of trial failure by selecting candidates that are likely to succeed and have already resulted in the identification of approved medications and promising candidates.[3,4]

A large number of computational DR approaches attempt to identify drug-disease associa-

tions by utilizing molecule structure, common pathways, or other known biological properties.[5] Signature matching and molecular docking use structural and chemical properties of molecules to identify similar drugs and therapeutic targets.[6] Other approaches use genome-wide summary statistics or biological pathway information to identify causal genes and new potential targets.[7] These methods generally attempt to use known information on the drug or disease in question to infer new treatment options.

Alternatively, electronic health records (EHRs) were used to identify potential alternative treatment targets based on documentation of side effects and clinical events.[8,9] Eguale et al.[10] used EHR data and Cox regression to associate off-label drug use with adverse drug events, Wu et al. have recently proposed another type of screening method using EHR records for the identification of drug-disease interactions.[11] Similarly, UK Biobank data has also been used to identify relations between treatment and phenotype, although these approaches generally focus on a specific phenotype and treatment pair. For instance, Ma et al.[12] used Cox regression in UK Biobank data to identify the benefits of glucosamine for type 2 diabetes. Pilling et al.[13] also used time-to-event modelling in UK Biobank to link lower vitamin D levels and hospitalization for delirium. Wu et al.[14] used PSM in UK Biobank for cost-benefit analysis of bariatric surgery.

Nevertheless, utilizing real-world data to isolate the effect of medication has proven challenging as this approach is highly prone to bias with the risk of creating spurious associations.[15] Indeed, in observational data, the characteristics of the treatment group are often very different from the average clinical study population. Propensity score matching (PSM)[16,17] is a statistical matching technique that attempts to associate subjects of the treatment group with similar subjects from the rest of the cohort to form a control group. Matched subjects have similar characteristics (as measured by selected covariates), limiting the impact of confounders in the estimation of the treatment effect. When time information of events is available, PSM can be combined with survival methods such as Cox regression[18] to estimate the relative risk between the treatment and control arms.[19]

In this work, we propose to model the risk of hospitalization w.r.t treatment for a large number of combinations of drugs and diseases. We effectively attempt to emulate thousands of clinical trials with hospitalization risk reduction as the endpoint. Our methodology is akin to genome-wide association studies (GWAS), in which a simple model is used to estimate the effect of a large number of loci in a hypothesis-free manner. As in GWAS, this form of drug-disease association study faces the risk of creating spurious relationships and requires further analysis, but can be seen as complementary to target-driven repurposing.[20] We applied our method to thousands of drug-disease pairs and showed that we can successfully re-identify associations that are already reported in UK Biobank, other observational cohorts, or controlled clinical trials. To the best of our knowledge, this is the first attempt to apply this type of systematic approach for treatment effect modelling.

## 2. Methods

### 2.1. *Dataset*

The UK Biobank[21] (UKBB) is a large observational dataset containing information on approximately 500K subjects over decades. During their initial visit to the UK Biobank assessment center, participants were interviewed about their medication use and completed a detailed questionnaire presenting questions on everyday habits, medical history, and mental health among others. A total of 1,233,630 treatments were reported, spanning 6745 different medications. Other biomarkers such as body mass index (BMI), blood pressure, and grip strength were also measured. Moreover, since the beginning of the study, more than 6 million hospitalization events were recorded in the form of an event date and a corresponding international classification of disease code (ICD10).

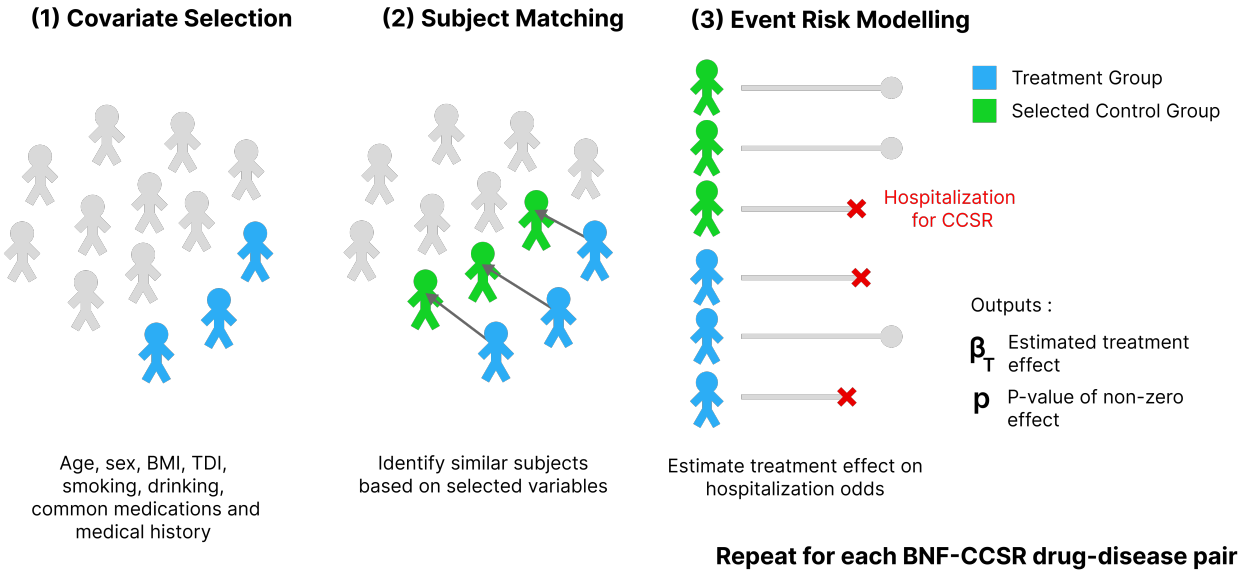


Fig. 1. Overview of the proposed method. It is composed of three main steps and is repeated for each BNF-CCSR drug-disease pair. (1) We use common comorbidities as covariates. Additionally, we included drugs and medical history, also respectively coded as BNF and CCSR, if they were present in more than 20% of the treatment group. We also remove subjects with a history of the CCSR code in question. (2) We use propensity score matching to find a similar non-treated subject for each subject of the treatment group, based on the selected covariates. (3) We use Cox regression, a proportional hazard ratio method, to estimate the treatment effect on hospitalization for the corresponding CCSR.

### 2.2. *Medication and Disease Selection*

ICD10 is a medical classification list from the World Health Organization used by many health organizations around the world that contains codes for over 70K diseases and symptoms. Some ICD10 codes represent similar phenotypes, for instance, I50.0, I50.1, and I50.9 correspond to congestive, left ventricular, and unspecified heart failure respectively. Using individual ICD10

codes for our analysis would be challenging due to the small number of events for each code, so we grouped them using Clinical Classifications Software Refined (CCSR)[22] v2023.1. CCSR is a classification system developed by the US Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project, that aggregates codes into clinically meaningful categories. We considered all CCSR categories spanning diseases of the circulatory system (CIR), endocrine, nutritional, and metabolic diseases (END) and symptoms, signs, and abnormal clinical and laboratory findings (SYM), as well as some others from diseases of the respiratory system (RSP), genitourinary system (GEN) and nervous system (NVS), totaling 77 phenotypes encompassing 3650 ICD10 codes.

Most of the medication types recorded in the UK Biobank dataset have very low frequency. Additionally, it is common for equivalent or similar active compounds to have different names, and no hierarchy is provided. To organize this data in a meaningful way, we mapped each medication to a corresponding British National Formulary (BNF) code. This code structure is used by the UK's National Health Service (NHS) to assign codes to drugs and chemicals and provides a fine-grained classification based on functionality. We used existing software[23] to map 3500 UKBB medications to 151 BNF codes. We only considered codes with at least 1000 subjects in the treatment group (power analysis with hazard ratio 0.6 and 80% power), resulting in 84 total BNF codes for analysis that include 93% of all reported medications in the UK Biobank during the first visit.

### 2.3. *Covariate and Subject Selection*

Medications (represented by BNF codes) and diseases (represented by CCSR codes) pairs were evaluated independently and the covariate and subject selection process was repeated for each pair. In total, we examined 6468 medication-disease pairs. We selected subjects from all available 500K participants who did not have a history of the CCSR code in question. Covariates can have a large impact on the estimated treatment effect and should be chosen carefully. In an attempt to be as general as possible, we used common demographics and risk factors: sex, age, BMI, Townsend deprivation index (TDI)[24] (related to poverty), smoking (current) and drinking habits (three times a week or more) as common covariates for all associations. For computational reasons, we capped the maximum number of subjects in the treatment group to 30,000, randomly sub-sampling when necessary.

To produce more precise matching and allow for more potential confounders, we also added medical history and medications as covariates if they were present in more than 20% of the treatment group. This percentage was evaluated for each individual drug-disease pair. Medical history was composed of both self-reported items and ICD10s prior to the first visit, grouped by CCSR coding. UKBB self-reported disease codes have their own representation, which were mapped to ICD10 and then to CCSR. Other medications were also selected by their corresponding BNF codes. This method of covariate selection has the advantage of being agnostic to the type of medication being considered. We used the same covariates for propensity score matching and Cox regression in all experiments.

### 2.4. *Propensity Score Matching and Pair Exclusion*

PSM consists in finding similar subjects in the control and treatment groups. This is done by fitting a logistic model and finding pairs of subjects that have the same probability of being in the same group. The unmatched subjects from the control group are then discarded. We used nearest neighbor distance as our matching method, and PSM was implemented in R using the matchit[25] package. PSM enables the estimation of the average effect of treatment in the treated individuals (ATT). In contrast to the average treatment effect (ATE), the ATT represents the effect of the drug on the treatment group, rather than the average population. As most drugs would not have any beneficial effect on a healthy population, we expect the effect of drugs for subjects that are already likely to be on treatment to be a more informative measure.

Despite still being widely used in retrospective studies, PSM has been criticized in the past[26] for potentially increasing imbalance between treatments and controls. However, this imbalance increase is only observed when groups are balanced initially, which was not the case in our experiments. Some alternatives to PSM, such as inverse probability of treatment weighting (IPTW) and Mahalanobis distance matching (MDM) were not considered due to their computational cost. In practice, we found PSM to produce balanced groups with minimal parameter tuning, and to be much more computationally efficient than other tested alternatives.

Additionally, unknown variables can bias the estimation of the treatment effect, to the point that the opposite effect can become statistically significant. This issue is not exclusive to PSM, and we observed that choosing the appropriate covariates was generally more impactful than the matching method itself. In some cases, the assignment of the treatment can deterministically depend on other variables, resulting in a lack of observation in the control group. Since PSM can introduce spurious relations between treatment and controls, careful interpretation of the treatment effects is always required. We report the number of balanced and unbalanced covariates for each pair in the summary statistics (mean standardized difference $< 0.1$).

We found that in some medication-disease pairs, some matched treatments and controls would be extremely dissimilar. Despite the large number of controls, it was simply not possible to match some subjects in the treatment groups in some cases. As an example, extremely morbidly obese subjects are almost always on the same medications. To address this issue, we computed the Huang distance[27] between each paired subject and discarded the pairs above an arbitrary threshold. The Huang distance was computed using both binary and normalized continuous covariates, treating CCSR history, sex, alcohol, and smoking habits as binary variables and the rest as continuous. In practice, we found only marginal improvements when excluding large-distance pairs.

### 2.5. *Cox Regression*

The Cox proportional hazard model[18] is a semi-parametric regression technique that estimates a relative hazard function, which represents a proportional risk of an event happening at time $t$.

The hazard function is of the form:

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)exp(\boldsymbol{X}_i \cdot \boldsymbol{\beta}) \tag{1}$$

where $\boldsymbol{X}_i$ represents the covariate vector for sample $i$, $\boldsymbol{\beta}$ are tunable regression coefficients also referred to as effect sizes and $\lambda_0$ is some common unknown hazard function that vanishes when estimating hazard ratios. Instead of binary categories, we considered a subject right-censored if no event had yet happened to that subject.

We fitted a separate model for each drug-disease pair. We considered all hospitalizations resulting in an ICD10 code contained in the CCSR category of choice as an event and only considered the first event if a subject had multiple events with the same CCSR code. Following the advice of Peter Austin,[19] we used a robust variance estimator and did not stratify on the matched sets. The output of the Cox regression is a treatment effect estimate $\beta_T$ and a corresponding P-value for the null hypothesis of a zero effect for the drug-disease pair. The Cox regression was implemented in R, using the survival[28] package.

Using only the information from the assessment center, we could not consider how long subjects had been on treatment, neither how long they would stay on it, nor the medication dosage. We also could not measure if subjects changed treatment over time. To attempt to minimize the impact of some of these limitations, we only considered events that happened before a given number of years after the assessment center visit and varied this time event window to 1, 3, 5, and 10 years. We also experimented with maximum pair Huang distances of 1, 2, 3, and no cut-off. Finally, we evaluated the impact of including common medical history and/or medications in the treatment group as covariates. A graphical overview of the method is presented in Figure 1.

## 3. Results

We applied our method to 77 disease categories and 84 medication types, resulting in 6468 potential drug-disease associations. Our results are reported in Figure 2. When comparing negative and positive associations, we observed a clear bias towards unfavorable effects ($\beta_T > 0$, corresponding to increased risk of hospitalization and hazard ratio greater than 1) for all parameters, although some configurations are less biased than others. Since these medications have been thoroughly tested for safety and side effects, we expect this ratio to be more balanced. We attribute this imbalance in significant associations to a failure to find appropriate matches in PSM, making the control group systematically healthier than the treatments.

We found that the bias towards unfavorable effects did not vanish when reducing the pair Huang distance cut-off, implying that this discrepancy is due to non-observable variables. When inspecting significant associations, we found that drugs that were already used as a treatment for a CCSR category were consistently associated with a higher hospitalization risk for the same CCSR. Our explanation for this observation is that treatment was prescribed to high-risk subjects without any hospitalization event or self-report, making the treatment group inherently more at risk than matched controls.

As an example, anti-diabetic drugs (BNF 6.1.2) were consistently associated with a higher risk of diabetes (CCSR END002). This is likely due to the fact that we could not match for prediabetes effectively, and thus the treatment group was much more likely to end up diabetic than

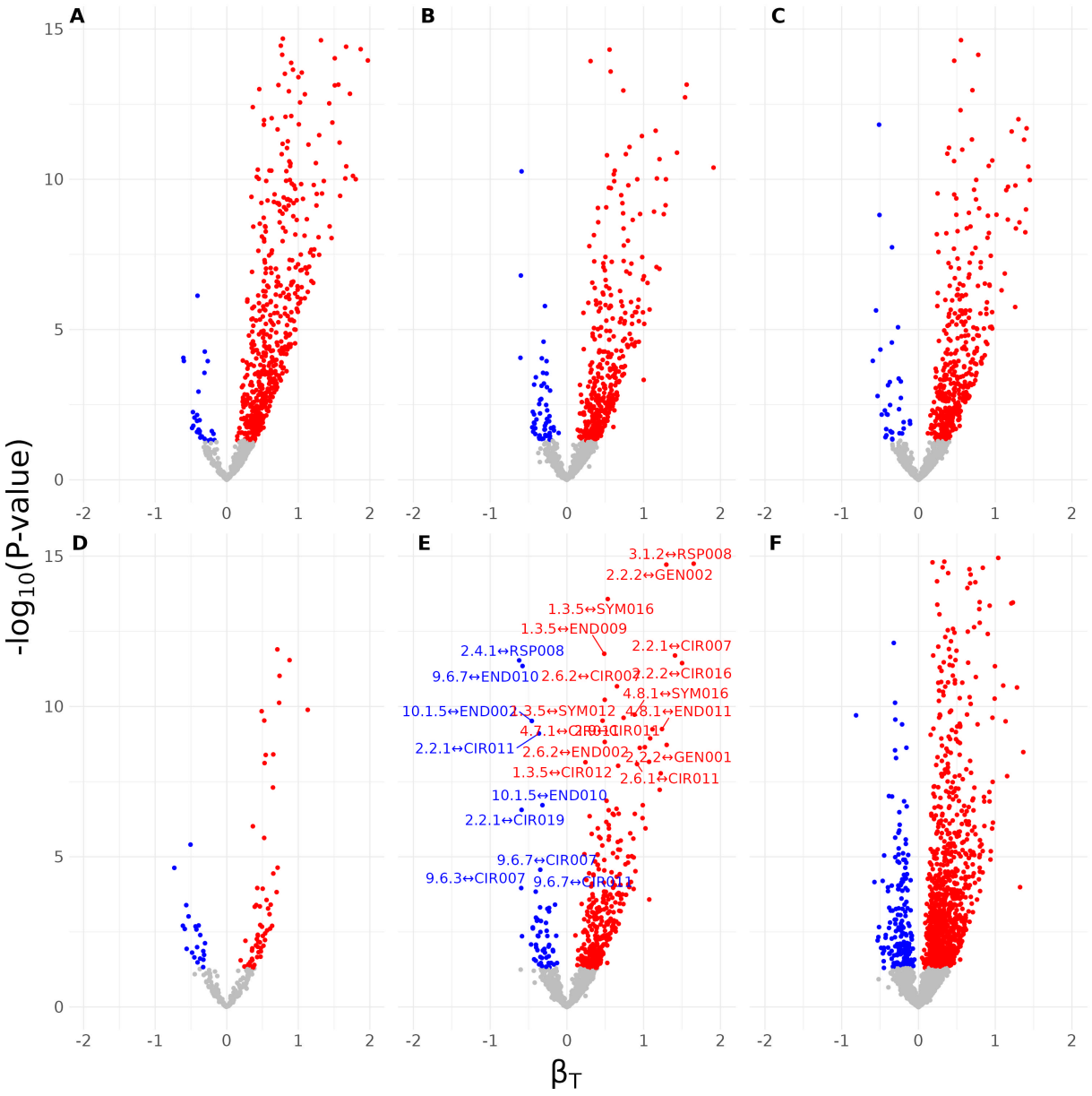


Fig. 2. Volcano plots of effect estimates for each drug-disease pair, spanning 84 medication categories and 77 phenotypes for multiple parameter choices. $\beta_T$: Cox regression coefficient; positive values indicate unfavorable effects (increased hazard ratio for hospitalization). Non-significant ($p > 0.05$) associations are reported in grey. Medications associated with a reduced or increased risk of hospitalization for the corresponding disease are reported in blue and red, respectively. **A**: Matching only with common covariates: sex, age, BMI, TDI, smoking, and drinking habits with an event time limit of 3 years. **B**: Matching with common covariates and medical history with an event time limit of 3 years. **C**: Matching with common covariates and other medications with an event time limit of 3 years. **D**: Matching with common covariates, medical history, and other medications with a maximum event time limit of 1 year post-visit. **E** : 3 year limit. **F** : 10 year limit.

the matched controls. Based on the previous results, we chose the combination of covariates, time, and Huang distance cut-offs that would result in the most balanced number of total associations (Maximum Huang distance of 3, Maximum time-to-event of 3 years, and including both medication and medical history). We used this configuration for all further analysis. Based on power analysis (60% hazard ratio and 80% power), we automatically discarded drug-disease pairs that included less than 100 events. We were able to estimate effects for 1013 pairs, and we used this number for correcting for multiple testing.

As we estimate the ATT, the correct interpretation of these measurements is that treated subjects would have a different risk of the corresponding CCSR code had they not taken the treatment, after correcting for all other known covariates. The root cause of this risk reduction cannot be inferred, and additional analysis is always required to determine the clinical relevance of this measured effect. Similarly, our method is also capable of measuring side effects that manifest as increased risks of hospitalization. We report all statistically significant effects after correcting for multiple testing (Bonferroni correction on the number of drug-disease pairs tested) in table 1, ordered by statistical significance. We expand further on each pairing in the next sub-sections.

Table 1. All statistically significant medications associated with a reduced risk of hospitalization for the corresponding disease ($p < 5 \cdot 10^{-5}$, after Bonferroni correction for multiple comparisons), ordered by P-value.

| BNF | CCSR | Medication | Disease | Hazard Ratio | P-value |
|---|---|---|---|---|---|
| 2.4.1 | RSP008 | Cardioselective Beta-blockers | COPD and bronchiectasis | 0.54 | 2.9e-12 |
| 9.6.7 | END010 | Multivitamins | Disorders of lipid metabolism | 0.56 | 4.5e-12 |
| 10.1.5 | END002 | Glucosamine | Diabetes mellitus | 0.63 | 3e-10 |
| 2.2.1 | CIR011 | Thiazides | Coronary atherosclerosis | 0.69 | 7.9e-10 |
| 10.1.5 | END010 | Glucosamine | Disorders of lipid metabolism | 0.73 | 1.9e-07 |
| 2.2.1 | CIR019 | Thiazides | Heart failure | 0.55 | 2.8e-07 |
| 9.6.7 | CIR007 | Multivitamins | Essential hypertension | 0.7 | 2.7e-05 |

### 3.1. *Cardioselective Beta-blockers and COPD*

Beta-adrenergic blocking agents or beta-blockers (BNF 2.4) used for COPD (CCSR RSP008) constitute one of our most significant positive drug-disease pairs ($\beta_T = -0.56, p = 10^{-10}$) with a corresponding hazard ratio for hospitalization risk of 0.57. Historically, the use of beta-blockers was discouraged for COPD as non-selective beta-blockers can reduce lung function.[29] Never-

theless, several retrospective observational studies have shown that usage of beta-blockers can reduce mortality and other exacerbations in COPD.[30,31]

The beta-blocker BNF encoding does not separate between cardioselective and non-selective compounds, so we split this category into two, 2.4.1 and 2.4.2 for cardioselective and non-cardioselective beta-blockers respectively. We found a stronger effect and smaller P-value ($\beta_T = -0.62$, $p = 2.9 \cdot 10^{-12}$) for category 2.4.1 w.r.t 2.4 while results of non-selective beta-blockers were not significant; this corroborates recent observational findings.[32,33] Thus, our results agree with the consensus that cardioselective beta-blockers are not only safe for patients at risk of COPD but could also reduce their risk of hospitalization.[34] The effect of cardioselective beta-blockers for patients with COPD is the subject of an ongoing phase IV clinical trial (NCT03566667).[35]

### 3.2. *Glucosamine and Diabetes Mellitus*

Glucosamine is a widely used supplement for osteoarthritis that is often taken daily and has anti-inflammatory properties. While glucosamine has been shown to induce insulin resistance in rodents[36] this effect does not appear to be present in humans.[37] Nevertheless, similarly to our findings, another recent UK Biobank study also showed the potential of glucosamine for the prevention of diabetes.[12] Since glucosamine does not impact blood sugar levels, glucose tolerance, or insulin resistance, this effect is likely not direct. However, there is an established relation between inflammation and the occurrence of diabetes,[38] and even support for inflammatory pathways to be involved in its pathogenesis.[39] The anti-inflammatory properties of glucosamine and the reduction of symptoms of arthritis might explain its apparent benefits for diabetes. We also found a reduced risk of hospitalization for disorders of lipid metabolism, a CCSR category that includes different types of hypercholesterolemia and hyperlipidemia (corresponding to ICD10 E78). Thus, long-term glucosamine supplementation might be beneficial for the prevention of diabetes and other metabolic diseases.

### 3.3. *Multi-vitamin Supplementation*

There is mixed evidence for the benefits of multi-vitamin (MVM) supplementation for general health,[40,41] with a general consensus from clinical trials that MVM supplementation does not reduce CVD mortality. Recently, Che et al.[42] found that multivitamin/mineral supplementation was associated with a modest reduction in CVD events in the UK Biobank. In contrast, we find that MVM supplementation is associated with a substantial reduction in risks of disorders of lipid metabolism and essential hypertension.

We suspect that the average MVM user in UK Biobank is more health-conscious than their matched counterparts or has had MVM and other supplementations for a long time before their visit to the assessment center, thus biasing our estimates. Subjects were matched for their history of hypertension, use of non-opioid analgesics, lipid-regulating drugs, and glucosamine in addition to the common covariates. Adding other confounding variables such as diet and exercise might reduce the estimated effect of MVM, although we leave this analysis for future research.

### 3.4. *Thiazides and Heart Failure*

We report that thiazides, a family of diuretics, are associated with a reduced risk of hospitalization for coronary atherosclerosis and heart failure. This coincides with the results reported in previous studies such as the SPRINT clinical trial,[43] which showed the importance of intense systolic blood pressure management for the risk reduction of cardiovascular events. Additionally, more than half of heart failure cases have a history of hypertension.[44]

When inspecting the treatments and matched controls, we found that only approximately 5% of the control group was on some form of non-thiazide diuretic. The proportion of other blood pressure medications such as beta-blockers were otherwise similar. 98% of the treatment group had a history of hypertension, while the ratio for the control group was 96%. It is possible that the observed reduction in hospitalization risk might generalize to other types of diuretics.

Interestingly, we observed an opposite effect for loop diuretics (LD, BNF 2.2.2) and heart failure. As the number of subjects on thiazides was significantly larger than the LD group, the LD treatment group was matched with a large proportion of subjects on some other diuretic, which was not the case for thiazides. Furthermore, LD are also more likely to be already used for the management of heart failure, thus biasing our estimates.

Recent studies support the use of thiazides for the treatment of heart failure. Using data from the SPRINT study, Tsjimoto et al.[45] found that thiazides decreased the risk of events for heart failure in non-diabetics. In the CLOROTIC trial[46] the combination of thiazides and LD proved to be effective for the treatment of acute heart failure. Unfortunately, only approximately one hundred subjects used both thiazides and LD in our dataset, making the estimation of the effect of the combination of both treatments unfeasible. Nevertheless, our results underline the importance of hypertension management for the prevention of heart failure and the potential of thiazide diuretics.

### 3.5. *Other Associations*

We found 92 statistically significant associations (after Bonferroni correction) for medications that increase the risk of hospitalization ($\beta_T > 0$, $p < 5 \cdot 10^{-5}$). Four medication types included 52 of these associations, all of which are reported in Table 2. As previously explained, some of these associations are known to be spurious, for instance, aspirin (BNF 4.7.1) does not cause an increased risk of hospitalization for hypertension (CIR007). However, since aspirin is commonly prescribed to individuals at risk of hypertension and other diseases it is associated with the phenotype in our analysis. We observe a similar effect for loop diuretics and multiple cardiovascular diseases.

We also observed that proton pump inhibitors (PPIs, BNF 1.3.5) were associated with an increased risk for 23 diseases. We offer three potential explanations. 1) Subjects on PPIs have systematically poorer health than their matched counterparts, either due to unknown variables or scarcity of suitable matches in the control group. 2) PPIs are used in the treatment of multiple diseases in the list or other related comorbidities, thus biasing our estimates. 3) PPIs have measurable side effects and increase the risk of hospitalization for multiple diseases. Since PPIs are used for gastric acid-related disorders and have several known side-effects,[47,48] it is plausible for some of these associations to be causal. Further analysis would be required

to estimate the causal effect of PPIs on these diseases, either by Mendelian randomization or a controlled study. We also come to a similar conclusion for anti-epileptic drugs (BNF 4.8.1), although the probability for these associations to be causal is lower.

Table 2. Medication associated with a higher risk of hospitalization ($\beta_T > 0$, $p < 5 \cdot 10^{-5}$) for the corresponding CCSRs, ordered by P-value from left to right. CIR: Diseases of the circulatory system; SYM: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; RSP: Diseases of the respiratory system; GEN: Diseases of the genitourinary system; END: Endocrine, nutritional and metabolic diseases

| **Drug Category** | **CCSRs** |
|---|---|
| Proton Pump Inhibitors | CIR007 ($p$ =3.4e-36), SYM006, END010, CIR011, RSP008, SYM016, END009, SYM012, CIR012, END007, GEN003, GEN002, END011, SYM001, CIR031 GEN001 , CIR026, SYM010, END002, SYM013, SYM014, SYM017, CIR016 |
| Loop Diuretics | GEN003 ($p$ =4e-22), CIR019, GEN002, CIR016, END011, CIR003, SYM016 GEN001, CIR015, CIR011, END002, CIR031 |
| Control of Epilepsy | SYM016 ($p$ =1.8e-10), END011, SYM010, RSP008, CIR007, GEN003, SYM012 SYM015, SYM001 |
| Non-opioid Analgesics | CIR007 ($p$ =1e-26), END010, CIR011, RSP008, END002, CIR012, CIR026, END009 |

We also found 58 other risk-lowering associations ($\beta_T < 0$, $p < 0.05$) that were not statistically significant after correcting for multiple testing. Several of these associations were also reported in the literature and could have potential clinical applications. Anti-diabetic drugs (BNF 6.1.2), mostly composed of metformin (86% of treated subjects) and blood sugar lowering medications, were associated with reduced risk of 7 disease categories including conduction disorders ($p = 0.0023$), cardiac dysrhythmias ($p = 0.034$) and heart failure ($p = 0.047$). This corroborates the known cardiovascular benefits of metformin and other anti-diabetic drugs.[49–51]

## 4. Discussion

In this work, we proposed a purely phenotypic screening approach for drug repurposing that consists in systematically measuring medication effects on hospitalization risk from observational data. We showed that we could re-identify known repurposing candidates using simple extensively tested techniques for causal inference and time-to-event modelling. Grouping drugs and diseases by functionality allowed us to gather enough events to estimate potential effects while keeping fine-grained categories. We estimated the risk of hospitalization, making our method inherently preventive although some results could generalize to already hospitalized patients. While our results mostly corroborate known associations, the data for this study has been available for ten years and this method can be applied to new cohorts and treatments.

Due to the nature of the examined data, our study presents multiple limitations. The generally low frequency of events for each CCSR code made the estimation of most effects impossible. While more events could have been included by increasing the time event limit, this would have also introduced more spurious associations. Without utilizing general provider longitudinal data, we could not estimate the approximate dosage, length of treatment, or

whether subjects swapped treatments after the first visit and we found a maximum time from visit of 3 years to be a good compromise. Medications were self-reported and no corresponding indication was provided. While matching for common medication has shown to produce less imbalance in general, it can also be counterproductive in cases where a single drug is used for multiple purposes and can result in inadequate matching. The quality of the matching itself is difficult to quantify as most of the bias comes from unmeasured variables, or due to irreconcilable differences between control and treatment groups.

Despite these limitations, biobanks have multiple advantages over typical EHR datasets. 1) All measurements were taken with the same methodology by a small number of assessment centers. 2) Measures such as BMI were taken at a single time point, making time-to-event analysis straightforward. In contrast, EHRs typically have a large portion of missing variables and information is spread over multiple records. 3) Subjects directly described in detail their medication intake and medical history. These variables would be more challenging to recover with EHR data and would likely be incomplete, as the subject history must be stitched up from past events. UK Biobank data allowed us to perform time-to-event analysis with relatively little pre-processing, and scaling up to thousands of tests was also straightforward to implement. To the best of our knowledge, we are the first to report associations for cardioselective beta-blockers, thiazides, and proton pump inhibitors in the UK Biobank.

Large-scale biobank data are a precious resource for understanding human health. While retrospective analysis is always biased and incomplete, it can be an effective tool to guide the design of future experiments that is complementary to other DR methods. Our proposed approach is especially effective at identifying repurposing candidates for preventive care of high-risk subjects. In the future, we plan on using longitudinal general provider prescription data to refine our estimates.

## 5. Code and Data Availability

Code used for the analysis and summary statistics for all drug-disease pairs in this manuscript is provided on a dedicated GitLab repository `https://gitlab.com/CGeorgantasCHUV/SYESTE`.

## 6. Acknowledgements

This research has been conducted using the UK Biobank resource under application number 80108, with funding from the Swiss National Science Foundation (Sinergia CRSII5_202276/1).

## References

1. N. Nosengo, Can you teach old drugs new tricks?, *Nature* **534**, 314 (June 2016).
2. N. Krishnamurthy, A. A. Grimshaw, S. A. Axson, S. H. Choe and J. E. Miller, Drug repurposing: a systematic review on root causes, barriers and facilitators, *BMC Health Services Research* **22**, p. 970 (July 2022).
3. S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla and M. Pirmohamed, Drug repurposing: progress, challenges and recommendations, *Nature Reviews Drug Discovery* **18**, 41 (January 2019).

4. V. Parvathaneni, N. S. Kulkarni, A. Muth and V. Gupta, Drug repurposing: a promising tool to accelerate the drug discovery process, *Drug Discovery Today* **24**, 2076 (October 2019).
5. J. G. Moffat, F. Vincent, J. A. Lee, J. Eder and M. Prunotto, Opportunities and challenges in phenotypic drug discovery: an industry perspective, *Nature Reviews Drug Discovery* **16**, 531 (August 2017).
6. L. Pinzi and G. Rastelli, Molecular Docking: Shifting Paradigms in Drug Discovery, *International Journal of Molecular Sciences* **20**, p. 4331 (January 2019).
7. W. R. Reay and M. J. Cairns, Advancing the use of genome-wide association studies for drug repurposing, *Nature Reviews Genetics* **22**, 658 (October 2021).
8. H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, M. Levy, A. Shah, X. Han, X. Ruan, M. Jiang, Y. Li, J. S. Julien, J. Warner, C. Friedman, D. M. Roden and J. C. Denny, Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality, *Journal of the American Medical Informatics Association: JAMIA* **22**, 179 (January 2015).
9. M. Zhou, Q. Wang, C. Zheng, A. John Rush, N. D. Volkow and R. Xu, Drug repurposing for opioid use disorders: integration of computational prediction, clinical corroboration, and mechanism of action analyses, *Molecular Psychiatry* **26**, 5286 (September 2021).
10. T. Eguale, D. L. Buckeridge, A. Verma, N. E. Winslade, A. Benedetti, J. A. Hanley and R. Tamblyn, Association of Off-label Drug Use and Adverse Drug Events in an Adult Population, *JAMA internal medicine* **176**, 55 (January 2016).
11. P. Wu, S. D. Nelson, J. Zhao, C. A. Stone, Q. Feng, Q. Chen, E. A. Larson, B. Li, N. J. Cox, C. M. Stein, E. J. Phillips, D. M. Roden, J. C. Denny and W.-Q. Wei, DDIWAS: High-throughput electronic health record-based screening of drug-drug interactions, *Journal of the American Medical Informatics Association: JAMIA* **28**, 1421 (July 2021).
12. H. Ma, X. Li, T. Zhou, D. Sun, Z. Liang, Y. Li, Y. Heianza and L. Qi, Glucosamine Use, Inflammation, and Genetic Susceptibility, and Incidence of Type 2 Diabetes: A Prospective Study in UK Biobank, *Diabetes Care* **43**, 719 (April 2020).
13. L. C. Pilling, L. C. Jones, J. A. H. Masoli, J. Delgado, J. L. Atkins, J. Bowden, R. H. Fortinsky, G. A. Kuchel and D. Melzer, Low Vitamin D Levels and Risk of Incident Delirium in 351,000 Older UK Biobank Participants, *Journal of the American Geriatrics Society* **69**, 365 (February 2021).
14. T. Wu, K. B. Pouwels, R. Welbourn, S. Wordsworth, S. Kent and C. K. H. Wong, Does bariatric surgery reduce future hospital costs? A propensity score-matched analysis using UK Biobank Study data, *International Journal of Obesity* **45**, 2205 (October 2021).
15. Drug repurposing using real-world data, *Drug Discovery Today* **28**, p. 103422 (January 2023).
16. P. Rosenbaum and D. Rubin, The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41 (April 1983).
17. U. Benedetto, S. J. Head, G. D. Angelini and E. H. Blackstone, Statistical primer: propensity score matching and its alternatives, *European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery* **53**, 1112 (June 2018).
18. D. R. Cox, Regression Models and Life-Tables, *Journal of the Royal Statistical Society. Series B (Methodological)* **34**, 187 (1972).
19. P. C. Austin, The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments, *Statistics in Medicine* **33**, 1242 (March 2014).
20. A. G. Reaume, Drug repurposing through nonhypothesis driven phenotypic screening, *Drug Discovery Today: Therapeutic Strategies* **8**, 85 (December 2011).
21. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman

and R. Collins, UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age, *PLOS Medicine* **12**, p. e1001779 (March 2015).
22. Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses.
23. A. Phil, UK Biobank Self Reported Medication Data parsing and matching.
24. B. Jarman, P. Townsend and V. Carstairs, Deprivation indices., *BMJ : British Medical Journal* **303**, p. 523 (August 1991).
25. D. Ho, K. Imai, G. King, E. Stuart, A. Whitworth and N. Greifer, MatchIt: Nonparametric Preprocessing for Parametric Causal Inference (June 2023).
26. G. King and R. Nielsen, Why Propensity Scores Should Not Be Used for Matching, *Political Analysis* **27**, 435 (October 2019).
27. Z. Huang, Clustering Large Data Sets With Mixed Numeric And Categorical Values1997.
28. T. M. Therneau, T. L. o. S.->. p. a. R. m. until 2009), A. Elizabeth and C. Cynthia, survival: Survival Analysis (March 2023).
29. W. MacNee, Beta-Blockers in COPD — A Controversy Resolved?, *New England Journal of Medicine* **381**, 2367 (December 2019).
30. F. H. Rutten, N. P. A. Zuithoff, E. Hak, D. E. Grobbee and A. W. Hoes, Beta-blockers may reduce mortality and risk of exacerbations in patients with chronic obstructive pulmonary disease, *Archives of Internal Medicine* **170**, 880 (May 2010).
31. P. M. Short, S. I. W. Lipworth, D. H. J. Elder, S. Schembri and B. J. Lipworth, Effect of beta blockers in treatment of chronic obstructive pulmonary disease: a retrospective cohort study, *BMJ (Clinical research ed.)* **342**, p. d2549 (May 2011).
32. Y.-L. Yang, Z.-J. Xiang, J.-H. Yang, W.-J. Wang, Z.-C. Xu and R.-L. Xiang, Association of β-blocker use with survival and pulmonary function in patients with chronic obstructive pulmonary and cardiovascular disease: a systematic review and meta-analysis, *European Heart Journal* **41**, 4415 (December 2020).
33. C.-M. Chung, M.-S. Lin, S.-T. Chang, P.-C. Wang, T.-Y. Yang and Y.-S. Lin, Cardioselective Versus Nonselective β-Blockers After Myocardial Infarction in Adults With Chronic Obstructive Pulmonary Disease, *Mayo Clinic Proceedings* **97**, 531 (March 2022).
34. B. Lipworth, J. Wedzicha, G. Devereux, J. Vestbo and M. T. Dransfield, Beta-blockers in COPD: time for reappraisal, *The European Respiratory Journal* **48**, 880 (September 2016).
35. J. Sundh, A. Magnuson, S. Montgomery, P. Andell, G. Rindler, O. Fröbert, M. Przybyszewska, A. Blomberg, M. Widmark, A. Palm, W. Greger, J. Ellingsen, L. Råhlén, T. Kipper, M. Hasselgren, C. Smith, F. Delijaj, K. Possler, J. Nilsson, N. Stenersen, H. Nguyen, D. Curiac, L. E. G. W. Vanfleteren, L. Johansson, F. Sjöberg, M. Ekström, J. S. Berglund, A. Lökke and the BRONCHIOLE investigators, Beta-blockeRs tO patieNts with CHronIc Obstructive puLmonary diseasE (BRONCHIOLE) – Study protocol from a randomized controlled trial, *Trials* **21**, p. 123 (January 2020).
36. L. Rossetti, M. Hawkins, W. Chen, J. Gindi and N. Barzilai, In vivo glucosamine infusion induces insulin resistance in normoglycemic but not in hyperglycemic conscious rats., *The Journal of Clinical Investigation* **96**, 132 (July 1995).
37. R. Muniyappa, R. J. Karne, G. Hall, S. K. Crandon, J. A. Bronstein, M. R. Ver, G. L. Hortin and M. J. Quon, Oral glucosamine for 6 weeks at standard doses does not cause or worsen insulin resistance or endothelial dysfunction in lean or obese subjects, *Diabetes* **55**, 3142 (November 2006).
38. K. E. Wellen and G. S. Hotamisligil, Inflammation, stress, and diabetes, *The Journal of Clinical Investigation* **115**, 1111 (May 2005).
39. Z. Tian, J. Mclaughlin, A. Verma, H. Chinoy and A. H. Heald, The relationship between rheumatoid arthritis and diabetes mellitus: a systematic review and meta-analysis, *Cardiovascular En-*

*docrinology & Metabolism* **10**, 125 (February 2021).
40. H.-Y. Huang, B. Caballero, S. Chang, A. Alberg, R. Semba, C. Schneyer, R. F. Wilson, T.-Y. Cheng, G. Prokopowicz, G. J. Barnes, J. Vassy and E. B. Bass, Multivitamin/mineral supplements and prevention of chronic disease, *Evidence Report/Technology Assessment* , 1 (May 2006).
41. J. Kim, J. Choi, S. Y. Kwon, J. W. McEvoy, M. J. Blaha, R. S. Blumenthal, E. Guallar, D. Zhao and E. D. Michos, Association of Multivitamin and Mineral Supplementation and Risk of Cardiovascular Disease: A Systematic Review and Meta-Analysis, *Circulation. Cardiovascular Quality and Outcomes* **11**, p. e004224 (July 2018).
42. B. Che, C. Zhong, R. Zhang, M. Wang, Y. Zhang and L. Han, Multivitamin/mineral supplementation and the risk of cardiovascular disease: a large prospective study using UK Biobank data, *European Journal of Nutrition* **61**, 2909 (September 2022).
43. The SPRINT Research Group, A Randomized Trial of Intensive versus Standard Blood-Pressure Control, *New England Journal of Medicine* **373**, 2103 (November 2015).
44. G. C. Oh and H.-J. Cho, Blood pressure and heart failure, *Clinical Hypertension* **26**, p. 1 (January 2020).
45. T. Tsujimoto and H. Kajio, Thiazide Use and Decreased Risk of Heart Failure in Nondiabetic Patients Receiving Intensive Blood Pressure Treatment, *Hypertension* **76**, 432 (August 2020).
46. J. C. Trulls, J. L. Morales-Rull, J. Casado, M. Carrera-Izquierdo, M. Snchez-Marteles, A. Conde-Martel, M. F. Dvila-Ramos, P. Llcer, P. Salamanca-Bautista, J. Prez-Silvestre, M. N. Plasn, J. M. Cerqueiro, P. Gil, F. Formiga, L. Manzano and CLOROTIC trial investigators, Combining loop with thiazide diuretics for decompensated heart failure: the CLOROTIC trial, *European Heart Journal* **44**, 411 (February 2023).
47. A. J. Schoenfeld and D. Grady, Adverse Effects Associated With Proton Pump Inhibitors, *JAMA Internal Medicine* **176**, 172 (February 2016).
48. C. E. Aubert, M. R. Blum, V. Gastens, O. Dalleur, F. Vaillant, E. Jennings, D. Aujesky, W. Thompson, T. Kool, C. Kramers, W. Knol, D. O'Mahony and N. Rodondi, Prescribing, deprescribing and potential adverse effects of proton pump inhibitors in older patients with multimorbidity: an observational study, *Canadian Medical Association Open Access Journal* **11**, E170 (January 2023).
49. G. Rena and C. C. Lang, Repurposing Metformin for Cardiovascular Disease, *Circulation* **137**, 422 (January 2018).
50. F. Luo, A. Das, J. Chen, P. Wu, X. Li and Z. Fang, Metformin in patients with and without diabetes: a paradigm shift in cardiovascular disease management, *Cardiovascular Diabetology* **18**, p. 54 (April 2019).
51. W. M. C. Top, A. Kooy and C. D. A. Stehouwer, Metformin: A Narrative Review of Its Potential Benefits for Cardiovascular Disease, Cancer and Dementia, *Pharmaceuticals* **15**, p. 312 (March 2022).