Targeted inference to identify drug repositioning candidates in the Danish health registries

Alexander Wolfgang Jung[1,2], Ioannis Louloudis[1], Søren Brunak[1], Laust Hvas Mortensen[1,2]
[1] Novo Nordisk Foundation Center for Protein Research,
Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[2] Statistics Denmark, Copenhagen, Denmark

Abstract

Electronic health records can be used to track diagnoses and drug prescriptions in large heterogeneous populations over time. Coupled with recent advances in causal inference from observational data, these records offer new opportunities to emulate clinical trials and identify potential targets for drug repositioning. Here, we run a hypothesis generating cohort study of Danes aged 50 to 80 years from 2001 to 2015 (n = 2,512,380), covering a total of 23,371,354 years of observations. We examine prescription drugs at ATC level-4 and their effect on 9 major disease outcomes. Using Bayesian time-varying Cox regression and longitudinal minimum loss estimation, our analysis successfully reproduces known drug-disease associations from clinical trials, such as the reduction in the 3-year absolute risk of death associated with Statins (ATC:C10AA) -0.8% (95% CI =[-1.2%, -0.5%]) and -0.8% (95% CI =[-1.3%, -0.2%]) for females and males, respectively. Additionally, we discovered novel associations that suggest potential repositioning opportunities. For instance, Statins were associated with a reduction in the 3-year absolute risk of dementia by -0.3% (95% CI =[-0.5%, -0.1%]) for females and -0.2% (95% CI =[-0.4%, 0.1%]) for males. Furthermore, Biguanides (ATC:P01BB) stands out as a particularly interesting candidate with absolute risk reductions across various outcomes. In total, we identified 76 potential drug-disease pairs for further investigation. However, it should be stressed that the emulation of clinical trials here is solely of hypothesis generating nature and identified effects need to be corroborated with additional evidence, preferably from RTCs, as the risk of confounding by indication in this study is substantial. In summary, this study provides a large-scale screen of prescribed drugs and their effect on major debilitating disease in the Danish health registries. This provides an additional source of information that can be used in the search for possible repositioning candidates.

Keywords: Real-World Evidence, Drug repositioning, Targeted maximum likelihood estimation (TMLE)

1 Introduction

The development of new drugs is a time-consuming and expensive process with high rates of attrition often caused by efficacy- and safety-related failures [1–3]. Drug repurposing, re-utilizing approved drugs for target scopes other than the intended purpose, provides an intriguing proposition due to the reduced risk of adverse side effects given the prior assessment and evaluation for safety and dosing [4]. Repurposing propositions have largely been based on pharmacology and retrospective analysis with the most notable successes having been serendipitous like sildenafil for erectile dysfunction [5] or thalidomide for Multiple Myeloma [6] and Acute Myeloid Leukemia (AML) [7].

More systematic studies to produce testable hypotheses range from experimental approaches like binding assays and phenotypic screening to computational methods like genetic association studies, molecular docking, signature matching, pathway mapping, or the mining of Electronic Health Records (EHRs).

EHRs contribute a rich longitudinal and phenotypic data source, providing real-world evaluations of drug usage in large heterogeneous patient cohorts over prolonged time periods. The era of big data along with the development of new methods in the context of causality from observational data [8, 9], has opened new opportunities to leverage these data for novel insights by utilizing observational data to emulate hypothetical trials [10, 11].

These approaches have already informed clinical decision-making [12–14] and are particularly useful when a classical randomized clinical trial may not be feasible, due to time or ethical constraints. This was especially apparent during the SARS-CoV-2 pandemic when causality methods were used to evaluate the comparative effectiveness of the different vaccinations and booster campaigns in a rapidly evolving environment [15]. Additionally, conducting an observational study with a target trail in mind can help avoid certain statistical pitfalls like immortal time bias [16] and provide more robust effect estimates [17].

Here we make use of the Danish National Prescription Registry (DNPR) [18] and the Danish National Patient Registry (LPR) [19] combining information on all prescribed drugs dispensed through Danish community pharmacies and secondary care diagnoses across all of Denmark since 1995. We examine the joint contributions of an individual's drug usage and their comorbidities on the corresponding risk of onset for 9 extensively studied major disease outcomes (Dementia, Extrapyramidal disorders, Coronary vascular disease (CVD), Renal failure, Chronic obstructive pulmonary disease (COPD), Liver disease, Inflammatory bowel disease (IBD), Cancer and Death). A schematic representation of the study can be seen in Figure 1. The first step in the analytic approach is based on a Bayesian version of a time-varying Cox regression model as described previously [20]. This provides a preliminary evaluation of the multivariate effect size of the dispensed drug with the specified disease outcomes. As a second step, we use longitudinal minimum loss estimation (Ltmle) [21, 22], a doubly robust causal inference method, to obtain robust effect estimates. While we do use methods from causality they are applied in a generic way across most drugs and various outcomes, rather than explicitly emulating a specific target trail, therefore, estimates should not be considered causal.

This study is of a hypothesis generating nature and will provide a broad screen of dispensed drugs and their effects on selected major disease outcomes, ultimately providing a set of potential repurposing targets that could be corroborated with further evidence in the literature and potentially taken up for additional testing.

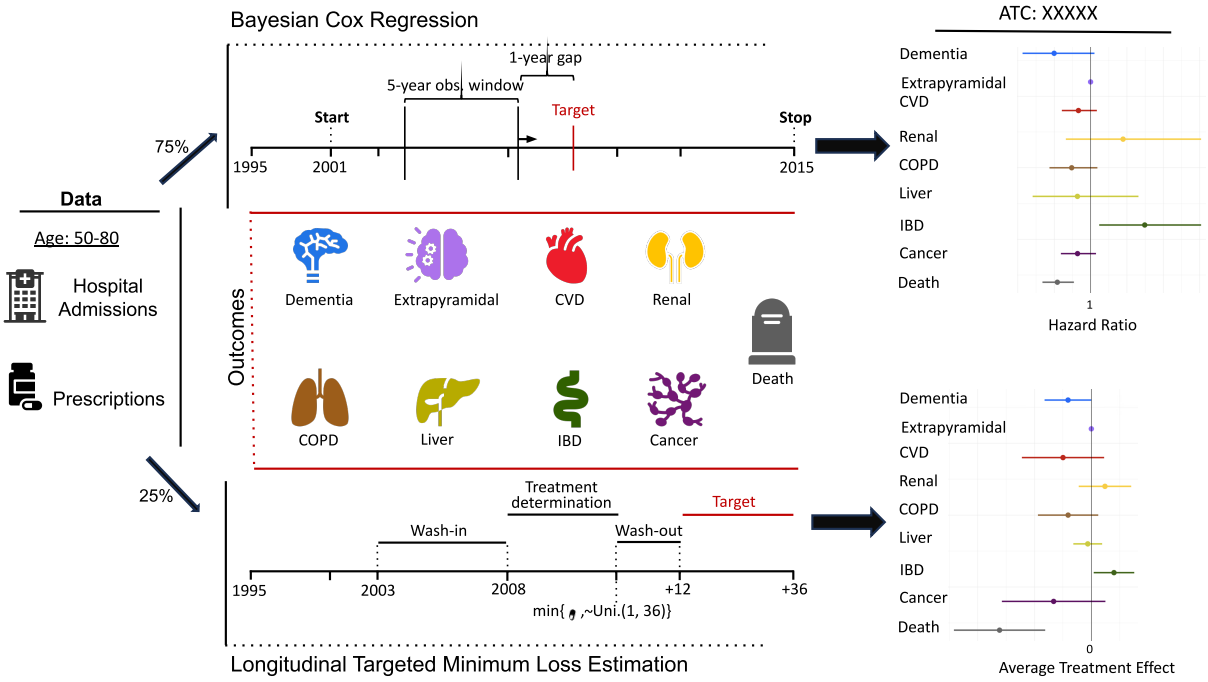

Figure 1 Schematic representation of the study. Information on secondary care admission and pharmacy dispensed drugs of individuals residing in Denmark aged 50-80 are collated. 75% of the individuals are used for an observational study design (upper part) and 25% are used for a target trail design (lower part). For either design, 9 outcomes are evaluated. The final results are effect estimates for a specific drug across the different outcomes, measured as either hazard ratio or average treatment effect for the different study designs, respectively.

2 Results

The study is based on all individuals residing in Denmark aged 50-80 years during a time window from 2001 until 2015 covering more than 2.5 million individuals and a combined 23 million years of observation containing 12 million diagnoses and 44 million dispensed drugs (Supplementary Table 1). At any given point in time, the covariates comprise binary indicators for secondary care diagnoses (ICD-10-3rd level codes e.g. E11 - Type 2 diabetes mellitus, Chapters: I-XVII: 1125/1034 - females/males) and binary indicators for dispensed drugs (ATC-4th level codes, e.g. A10BA, total: 472/458 - females/males) in the past 5 years. Separate models for females and males are estimated to investigate potential differential effects between the sexes.

In total we make use of two different study designs. First, one that follows standard observational studies containing a cross-section of the population (75%), on which we estimate a time-dependent Cox regression model for each of the 9 outcomes as a 1-year ahead prediction (upper section Figure 1). Second, one that uses an emulated target trail design covering the remaining 25% of the population to obtain robust estimates for all drug-disease pairs (lower section Figure 1). The full protocol for the emulated target trail can be found in the Methods section. In brief, the hypothetical trial for a specific treatment (ATC drug) and outcome starts in 2008. Every individual aged 50-80 who did not have the outcome yet and was not on treatment in the past 5 years is eligible to join. Each individual is assigned a random time between 1 and 36 months. If an individual starts treatment during this time window they are in the treatment arm otherwise they are assigned to the control arm. An individual's start time is either the allocated random time or the time when they start treatment, whichever comes earlier. An additional 12-month wash-out period is added to the start time to avoid confounding by indication but also to allow for a phase-in time of the drug. Subsequently, the next 36 months are used as the observation time on which the counterfactual estimates are based.

An overview of the number of individuals in the two designs as well as some basic characteristics can be seen in Table 1 (additional information can be found in Supplementary Table 1). The numbers for the emulated target trail are only approximate as they depend on each treatment/outcome pairing. A table for ATC:C10AA (Statins) and CVD is given in Supplementary Table 2 with all other combinations provided in the Supplementary Data.

Table 1 Data overview. Number of individuals in the two study designs split by sex with some basic characteristics. Additionally the number of events for each outcome is shown as well as the number of excluded individuals due to protocol violation e.g. prior disease indication. The numbers for the target trail are only approximate and should be understood as a rough guide as each case is dependent on the treatment/outcome combination.

	Bayesian Cox Regression				Emulated Target Trial			
	Female		**Male**		**Female**		**Male**	
Individuals:	901,833	51.29%	856,426	48.71%	218023	51.38%	206301	48.62%
Age:								
36-46	240,559	26.67%	237,831	27.77%	0	0.00%	0	0.00%
46-56	258,033	28.61%	258,120	30.14%	53,467	24.52%	53,114	24.36%
56-66	204,272	22.65%	199,519	23.30%	89,204	40.91%	88,261	40.48%
66-76	149,047	16.53%	126,268	14.74%	57,924	26.57%	51,579	23.66%
76-86	49,922	5.54%	34,688	4.05%	17,428	7.99%	13,347	6.12%
	Events:	**Exclude:**	**Events:**	**Exclude:**	**Events:**	**Exclude:**	**Events:**	**Exclude:**
Dementia	18,731	4,421	17,321	4,566	2119	1,993	1,842	1,966
Extrapyramidal	6,594	2,884	7,671	2,623	608	1,226	750	1,234
CVD	91,671	41,656	133,700	66,726	7,373	16,950	9,832	28,185
Renal	13,626	1,971	23,207	3,117	1,641	993	2,643	1,799
COPD	47,915	20,584	46,070	18,359	4,035	8,988	3,740	7,688
Liver	11,468	8,820	14,286	10,550	954	2,894	1,168	3,082
IBD	16,313	12,465	11,718	9,565	1,526	4,263	1,057	3,106
Cancer	94,895	57,604	105,562	28,932	7,946	20,435	9,780	12,247
Death	101,615	0	134,759	0	9,604	0	12,204	0

Risk assessment of comorbidities and medication history

To understand the overall contribution of secondary care diagnoses and dispensed drugs to the risk of developing one of the 9 outcomes, as well as to gain an initial estimate of effect sizes for different drugs, we conducted a classic observational study using penalized time-dependent Cox regressions.

Overall, as depicted in Figure 2a and Supplementary Table 3 age-adjusted concordance evaluated on an independent test set demonstrates good discrimination across most of the 9 outcomes with an average concordance of 0.692 (s.d.=0.08) and 0.68 (s.d.=0.079) for females and males, respectively. Cancers exhibit the least predictability with a concordance of 0.552 (95% CI =[0.548, 0.556]) for females and 0.56 (95% CI =[0.556, 0.564]) for males in line with previous results [23]. Conversely, renal failure and death show the best discrimination with 0.798 (95% CI =[0.79, 0.806]) and 0.798 (95% CI =[0.796, 0.80]) for females and 0.775 (95% CI =[0.769, 0.781]) and 0.774 (95% CI =[0.772, 0.776]) for males. Generally, discrimination between the sexes is similar, with the largest difference being observed for COPD with a concordance of 0.684 (95% CI =[0.68, 0.688]) for females and 0.62 (95% CI =[0.614, 0.626]) for males.

As the aim of this study is to identify potential candidates for repurposing, we focus on significant negative effect estimates based on the highest posterior density of 95% (HPD) [Figure 2b]. For females, there are a total of 207 drugs associated with the 9 outcomes. Extrapyramidal disorders show the fewest associations with only 4, while death has the most associations with 59. Similarly, for males, a total of 189 drugs are associated with the outcomes, with the fewest associations identified for extrapyramidal disorders and IBD, each with only 1 drug, and the most associations identified for death, with 51 drugs. Comparing across sexes, we identify a total of 98 drugs that show an effect in both, with extrapyramidal disorders having none, IBD and cancer having only 1 each, while death has the most with 33 common associations. Forest plots for all significant estimates, irrespective of the direction of the effect for each outcome, can be found in Supplementary Figures 1-9. The entire set of estimates can be found in the Supplementary Data.

Overall, effect estimates largely agree across sexes, showing high degrees of correlation between the log(hazard) estimates, as depicted in Figure 2c. All outcomes show Pearson correlations above 0.6, except for cancers, which shows a correlation of 0.459 [Supplementary Table 4]. Visual inspection of Figure 2c reveals that most estimates lie on the diagonal, indicating good agreement. However, some points lie on the respective axes, indicating estimates close to 0 for either sex. This does not necessarily reflect a true effect size of 0 but might instead be a result of the penalization term and a corresponding lack of power.

Further, summarizing the effects within the corresponding ATC chapters in Figure 2d reveals overall patterns of drug/disease pairs. For instance, drugs in the chapter Alimentary Tract and Metabolism (A) show an effect for all 9 outcomes, followed by drugs categorized in the Cardiovascular system (C) and drugs in Antiparasitic products, insecticides, and repellents (P), both missing associations with extrapyramidal disorders only. No negative associations are found for drugs in the chapter Antineoplastic and immunomodulating agents (L).

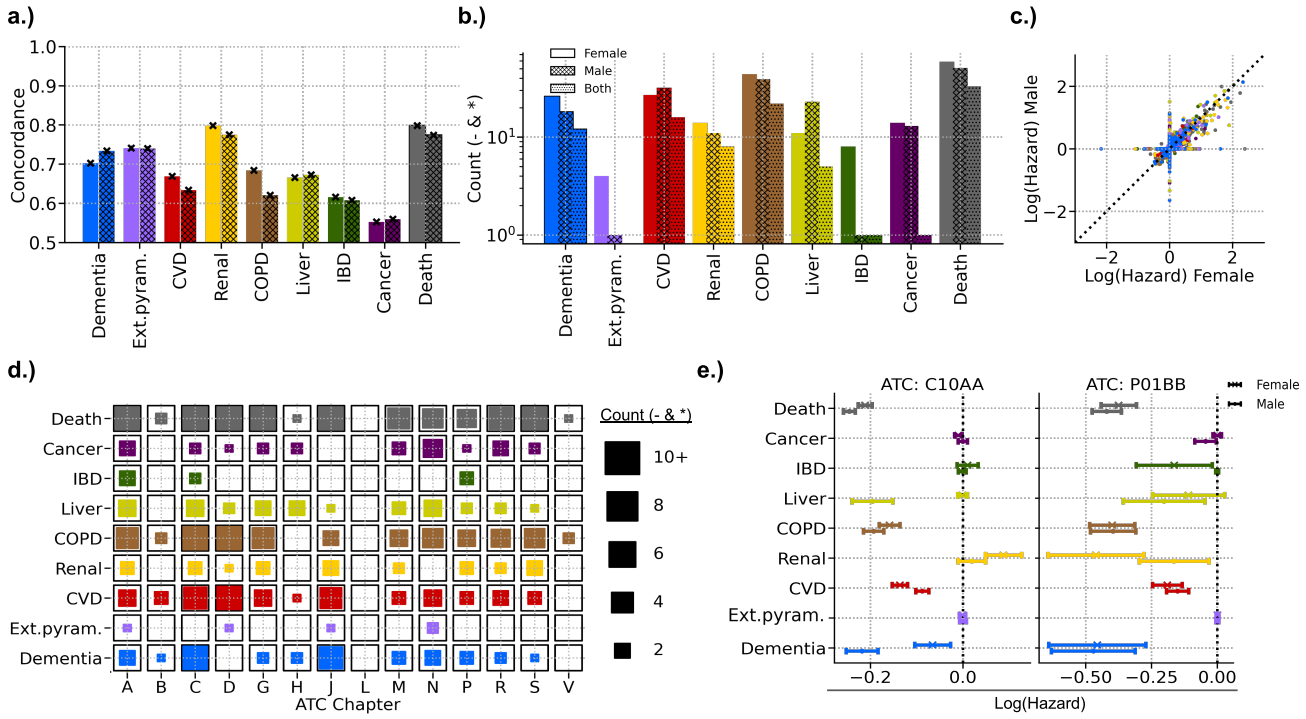

Figure 2 **a**: Age-sex adjusted concordance on test set (25%) across all outcomes. **b**: Number of negative and significant effects (based on highest posterior density 95%) across all outcomes split by sex and combined (counting effects that are present in both sexes). **c**: Scatter plot of the log(hazard) estimates between the model for females and males colored by the corresponding outcome that is estimated. **d**:Number of negative and significant associations aggregated by ATC chapter for each outcome. Effects for females and males are combined and treated as individual estimates. **A** ALIMENTARY TRACT AND METABOLISM, **B** BLOOD AND BLOOD FORMING ORGANS, **C** CARDIOVASCULAR SYSTEM, **D** DERMATOLOGICALS, **G** GENITO URINARY SYSTEM AND SEX HORMONES, **H** SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS, **J** ANTIINFECTIVES FOR SYSTEMIC USE, **L** ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS, **M** MUSCULO-SKELETAL SYSTEM, **N** NERVOUS SYSTEM, **P** ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS, **R** RESPIRATORY SYSTEM, **S** SENSORY ORGANS, **V** VARIOUS. **e**: Forest plot of the effect estimates for ATC:C10AA (Statins) and ATC:P01BB (Biguanides) across all outcomes and split by sex.

However, the potentially most meaningful insights can be gained when examining individual drugs across all outcomes simultaneously, as exemplified in the case of ATC:C10AA (Statins) and ATC:P01BB (Biguanides) in Figure 2e. Consistent with published clinical trial results [24, 25] ATC:C10AA shows a reduced risk of death, with a log(hazard) of -0.211 (95% CI =[-0.226, -0.196]) for females and -0.246 (95% CI =[-0.259, -0.233]) for males, as well as a reduced risk of CVD, with log(hazard) of -0.136 (95% CI =[-0.153, -0.12]) for females and -0.088 (95% CI =[-0.102, -0.074]) for males.

Surprisingly, ATC:P01BB, a Biguanide used for the treatment and prevention of Malaria shows clear negative associations across most of the evaluated disease outcomes, with similar performance across the sexes. However, this could very well be due to unmeasured confounding by indication, as people who use anti-malaria drugs might be traveling and hence are most likely in an overall healthy state. On the other hand side, Biguanides classified in ATC:A10AB, used in diabetic care e.g. Metformin, show either no effect or an increase in risk [Supplementary Figure 10].

Pseudo-causal evaluation of drug-disease pairs

To further gain insights into the reliability of the estimates, we perform a pseudo-causality analysis across most drug/disease pairs (only combinations with at least 1000 treated individuals). As mentioned earlier, this analysis is conducted in a generic way to scale to the number of combinations analyzed here, a total of 890 for females and 742 for males, rather than through a carefully crafted target trial; therefore, it is termed pseudo. While we do control for a large extent of an individual's medical history, there might be imbalances between the groups compared for unobserved confounders, and estimates should be interpreted with caution.

In total, we identify 76 drugs that show a significant negative association with the outcomes, with 52 for females, 24 for males, and 12 common across sexes. Dementia and death exhibit the most associations, with 21 and 10 for females, and 4 and 10 for males, respectively. Cancer and IBD display the fewest associations for females, with only 1 drug identified for each, while IBD shows no effects for males. Common associations between sexes are only identified for dementia, CVD, COPD, and death [Figure 2a].

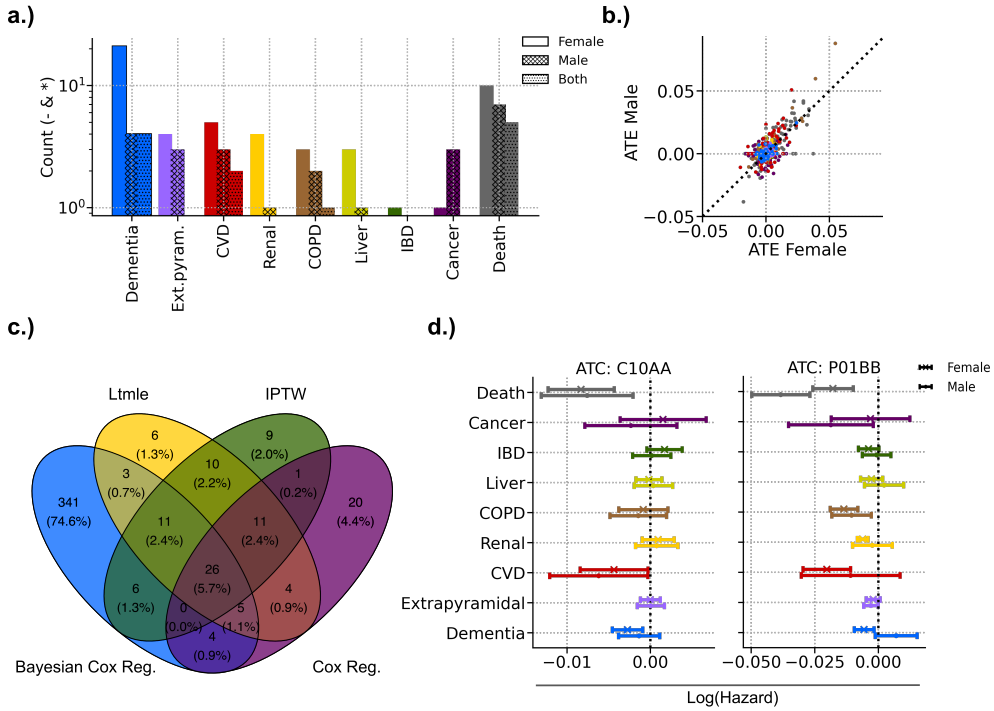

Figure 3 **a**: Number of negative and significant effects (based the 95% confidence interval) across all outcomes split by sex and combined (counting effects that are present in both sexes). **b**:Scatter plot of the average treatment effect (ATE) estimates between the model for females and males colored by the corresponding outcome estimated. **c**: Venn diagram of the count of overlapping negative and significant effects between the observational study and estimators based on the emulated target trial, including the longitudinal target minimum loss estimator (Ltmle), inverse probability of treatment weighted estimator (IPTW), and a simple Cox regression. **d**: Forest plot of the effect estimates for ATC:C10AA (Statins) and ATC:P01BB (Biguanides) across all outcomes and split by sex.

Overall, estimates between females and males appear similar, as can be seen in Figure 2b, with most estimates being close to the diagonal. This similarity is also reflected in the correlation of the average treatment effect (ATE) estimates between the sexes, with all outcomes showing a correlation of at least 0.2, except for cancers, which have a correlation of 0.129 (Supplementary Table 5).

Further, estimates are relatively stable between the two study designs and across different approaches, as illustrated in Figure 2c. Here, we present a Venn diagram of the identified significant negative associations for drug-disease pairs aggregated over the sexes. We compare the estimates from the observational Bayesian Cox regressions with the Ltmle estimates, inverse probability of treatment weighting estimates (IPTW), and estimates from simple Cox regressions fitted to the emulated trial data. A total of 45 (9.9%) estimates appear similar between the Bayesian Cox regression and Ltmle. A large fraction of 341 (74.6%) associations are only identified in the Bayesian Cox regressions, however, this is also expected as this design has the most power and is potentially more prone to identifying spurious relations. Estimates largely overlap with a total of 26 (5.7%) associations identified in all approaches. A table showing all estimates across the approaches can be found in the Supplementary Data.

Lastly, we can once again examine the effects of individual drugs across all outcomes simultaneously, as exemplified here by ATC:C10AA (Statins) and ATC:P01BB (Biguanides) in Figure 3d. Several of the effects identified for ATC:C10AA in the observational design vanish, with the main known effects from clinical trials remaining significant albeit slightly attenuated. Death shows an absolute risk reduction of -0.008 (95% CI =[-0.012, -0.005]) and -0.008 (95% CI =[-0.013, -0.002]) over a 3-year period for females and males, respectively, while CVD indicates a reduced absolute risk of -0.005 (95% CI =[-0.008, -0.0]) and -0.006 (95% CI =[-0.012, -0.0]).

Interestingly, we still observe a significant effect for ATC:C10AA and dementia in females, with a reduced absolute risk of -0.003 (95% CI =[-0.005, -0.001]). ATC:C10AA also shows a negative effect for males, with an absolute risk reduction of -0.002 (95% CI =[-0.004, 0.001]), although there is no clear evidence for a sign effect. While a potential link between statins and dementia risk has been proposed earlier [26–28], evidence so far is inconclusive [29] and further investigation may be warranted.

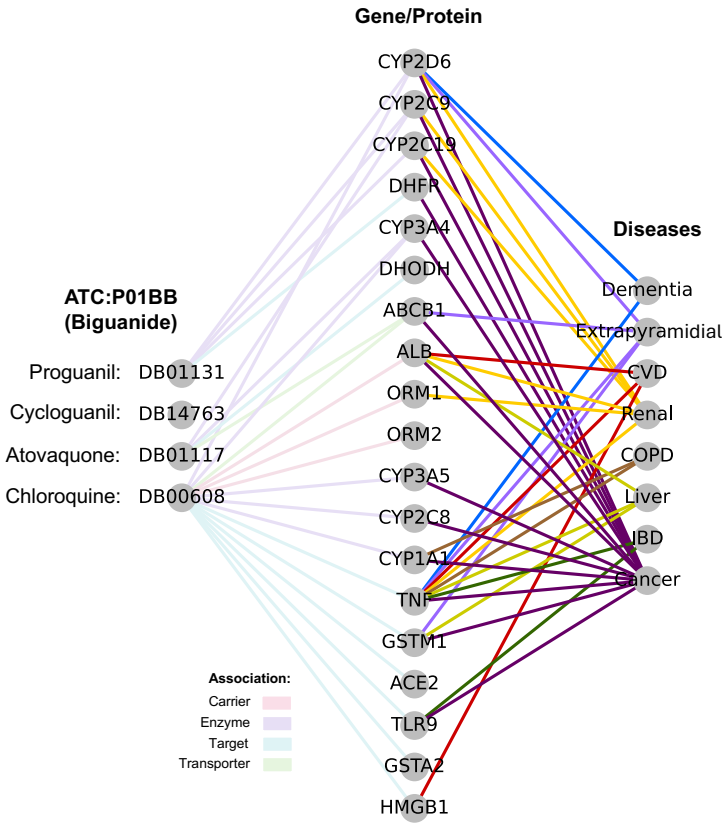

Figure 4 PrimeKG knowledge graph extract for drugs in ATC:P01BB (Biguanides) and their association with genes/proteins and their respective associations with the disease outcomes as of June 2024.

Surprisingly, many of the effects identified in the observational study for ATC:P01BB persist in the emulated trial design. In total, we find 8 significant effects for ATC:P01BB across all outcomes and both sexes. Death shows a clear 3-year absolute risk reduction of -0.0178 (95% CI =[-0.0258, -0.01]) and -0.0384 (95% CI =[-0.0498, -0.027]) for females and males, respectively. Further, we identify a potential absolute risk reduction for cancers of -0.003 (95% CI =[-0.0186, 0.0124]) in females and a significant reduction in absolute risk of -0.017 (95% CI =[-0.0354, -0.002]) for males. Most of the effects are similar between the sexes, albeit with stronger evidence of a sign effect in females. Dementia is the only exception to this, with a significant absolute risk reduction of -0.0056 (95% CI =[-0.01, -0.002]) in females but a potential absolute risk increase of 0.007 (95% CI =[-0.001, 0.0152]) in males.

Looking at ATC:A10AB in Supplementary Figure 10, a different Biguanide containing Metformin, we mostly see no clear effects, with the exception of death in males with an absolute risk increase of 0.0155 (95% CI =[0.002, 0.0289]) and liver diseases in males with an absolute risk increase of 0.006 (95% CI =[0.001, 0.0122]).

While the effects identified for ATC:P01BB appear interesting as a potential drug for further investigation, as we have cautioned in the previous section, there are potential mechanisms of confounding that we cannot control for, which could have biased the effects towards a reduction in risk.

The purpose of the study is solely of hypothesis generating nature and hence the effects need to be additionally verified in other more targeted studies and further corroborated through additional evidence. A minimal next step would be to investigate biomedical databases for potential links. As an example we looked at links between drugs in ATC:P01BB and our disease outcomes in the precision medicine database PrimeKG [30]. Traversing the knowledge graph from associations of associated drugs to genes/proteins as either carrier, enzyme, target or transporter to subsequently the association of the genes/proteins to our disease outcomes, as is shown in Figure 4, reveals a possible relation to several of the outcomes, with Chloroquine as a particularly outstanding case. Another possible mechanism of action reported in the literature is the link to the use of Hydroxychloroquine in the treatment of rheumatoid arthritis and other inflammatory rheumatic diseases [31]. Generally, we do see several anti-inflammatory related drugs e.g. Corticosterioids (ATC: C05AA, D07AB, H02AB, R01AD) or Anti-inflammatory and anti-rheumatic agents, non-steroids (ATC: M01AB, M01AX) with potential repositioning effects, indicating possibly underlying inflammatory aspects to some of the disease outcomes (Supplementary Figure 11). Further, recent studies suggest potential positive effect of Hydroxychloroquine on dementia risk [32] or Atovaquone in treatment of non-small cell lung cancer [33]

3 Discussion

Overall, this study demonstrates that EHRs, when combined with methods from causality, may be helpful in identifying novel associations that warrant further investigation. We conducted a comprehensive screening of most ATC level-4 drugs across a range of disease outcomes. All estimates are provided and can serve as a foundation for subsequent research or as supplementary evidence to support findings obtained from alternative approaches.

The most effective way to utilize this type of data and methods is through a carefully crafted emulated target trial with proper inclusion criteria. However, this is only possible with a specific hypothesis in mind. While, our approach can easily be applied in a generic way across most combination, this also makes it more prone to misspecification and hence should be evaluated carefully and understood more from an explanatory viewpoint.

Other limitations of our approach include the possibility of confounding from various sources. While we do control for a comprehensive set of medical information on individuals, there are certain aspects that may not be adequately addressed. Improved access to a broader range of clinical data could potentially mitigate some of these issues; however, the effects should be assessed with appropriate domain expertise to identify potential biases and directions of influence. Further, information on medications only contains prescriptions and does not necessarily reflect actual use. These two sources could explain the effects found for ATC:P01BB. First, it is not clear if individuals actually are taking the drug or only received a prescription due to planned travels. Second, individuals that are going on long-distance travels are probably healthier and potentially socioeconomically better suited for which we can only partially control.

A potential extension to our approach could involve incorporating multiple time-points for the emulated trial, thereby enhancing the overall power of the approach. Additionally, this approach would enable the study of outcomes that are typically rare, potentially offering more opportunities for drug repositioning. However, technical issues arise as then individuals could be part of multiple trials.

Furthermore, our design results in a real-time shift between the treatment and control arms, where treatment consistently occurs before the allocated random time. This shift should not significantly impact effect estimates unless there is a noticeable change in incidence within a small time window. However, during rapid shifts in disease incidence, such as those observed during the SARS-CoV-2 pandemic, this could become important and should be

taken into consideration. We conducted evaluations of the potentially introduced bias through simulations (see supplementary data) and found no measurable effect.

Finally, our approach assesses a wide array of combinations. While we utilize shrinkage priors and incorporate multiple cohort splits (such as data and sex splits), thereby offering multiple lines of evidence, we do not adjust for multiple hypotheses. It is important to emphasize that the objective of the study is not to make inferential statements about a particular effect, but rather to explore and screen for new targets.

4 Methods

Observational study

Data Sources: This study retrospectively utilizes data from the Danish health registries, which include the Central Person Registry (CPR), the Danish National Patient Registry (LPR), the Death Registry (DR), and the Danish National Prescription Registry (DNPR). Individuals born in or residing in Denmark for more than 3 months are registered. All registries have been linked via a unique personal identifier. Data compilation spans from January 1, 1995, to December 31, 2014.

Cohort: We included all individuals aged 50 to 80 who were alive on January 1, 2001, and who had been residing in Denmark continuously since at least January 1, 1995. This inclusion criterion ensures that all participants have a minimum of five years of recorded medical information at any given point in time. Participants exited the cohort upon reaching the age of 80 or due to exclusion criteria such as emigration, end of follow up or death, whichever occurred first. Individuals who emigrated after January 1, 2001, are censored at the point of emigration and remain so for the duration of the study. Individuals with events of interest occurring prior to January 1, 2001, are excluded from the study. The primary observational period for model fitting and evaluation extends from January 1, 2001, to December 31, 2014. The cohort is divided into three subsets: (i) a training set (70%), utilized for model training and development; (ii) a validation set (5%), employed for initial model evaluation and to determine the optimal penalization strength; (iii) a test set (25%), used for the final model assessment.

Covariates: The covariates include binary indicators for secondary care diagnoses extracted from the Danish National Patient Registry (LPR), utilizing ICD-10 codes up to the third level of specificity, recorded across chapters I-XVII (e.g., E11 for Type 2 diabetes mellitus). Considering the gender specificity of some diagnoses, we filtered indicators relevant to each sex, resulting in a total of 1,125 indicators for females and 1,034 indicators for males. Moreover, we limited the indicators to records from the preceding five years at any given point in time to capture recent health changes. Similarly, binary indicators for dispensed medications, recorded in the Danish National Prescription Registry (DNPR), are included. These medications are classified using the ATC system at the fourth level (e.g., A10BA for Biguanides). Due to the existence of sex-specific medications, we identified a total of 472 indicators for females and 458 for males, respectively. These indicators reflect medication usage in the past five years.

Outcomes: We consider a total of 9 outcomes including Dementia (ICD10: F00-03, G30-31), Extrapyramidal disorders (ICD-10: G20-26), Coronary vascular disease (CVD) (ICD-10: I21-26, I46, I50, I60-64), Renal failure (ICD-10: N17-19), Chronic obstructive pulmonary disease (COPD) (ICD-10: J41-J44, J47), Liver disease (ICD-10: K70-77), Inflammatory bowel disease (IBD) (ICD-10: K50-52), Cancers (ICD-10: C00-96, D37-48 excluding: C44, D45) and Death.

Statistical analysis: We fit time-dependent Bayesian Cox models with shrinkage priors for each of the nine outcomes and for each sex, using age as the underlying timeline. These models are solely fitted on the training set. Individuals are considered at risk upon reaching the inclusion age or the age at which they enter the cohort, whichever comes first. They are followed until the occurrence of a specific outcome, death, emigration, or the end of the follow-up period. Covariates are treated as time-dependent consisting of binary indicators for diseases and medications within the preceding five years. The effects of these covariates are modeled through a linear predictor. To prevent the inclusion of data that may only reflect the diagnostic process leading up to an outcome, we introduce a one-year gap between the occurrence of an event and its associated covariates. This approach ensures that evaluations are based on predictions made at least one year in advance. For additional details on the method and its implementation, we refer to Jung et al. (2022, 2023)[20, 23]. The final model specification uses a Student-T distribution with location=0, scale=0.001, and 1 degrees of freedom as the prior. A lower-rank(50) Multivariate-Normal distribution is used as the distributional family for stochastic variational inference. We perform stochastic gradient

descent updates using batches of 8196 randomly selected individuals. Confidence regions or highest posterior densities (HPD) are determined based on the posterior distributions, typically covering a 95% confidence interval unless stated otherwise. The concordance index serves as the primary metric for evaluating the fits of the models.

Emulated target trial

Data Sources: We utilize the same dataset as in the observational study, specifically, all individuals from the test set. As the test set has only been used to evaluate the concordance index for the observational study, it constitutes an independent data subset for estimation purposes. In principle, all individuals from the test set are included; however, additional restrictions will apply based on the targeted trial design, which we address below.

Covariates: The covariates utilized for estimation mirror those employed in the observational study. However, instead of treating them as time-dependent variables, we focus on a single time point: the start time of the emulated trial, excluding the washout period. From this time point, we construct binary indicators representing medication usage and acquired diseases over the past five years. Additionally, we apply a filtering criterion to the covariates, ensuring a minimum frequency of 0.01 in either the entire population, the treatment group, the untreated group, the event group, the non-event group, or any combination thereof, for each treatment and outcome pairing separately. This step aims to eliminate covariates that occur in only a small fraction of individuals across all possible subgroups, thereby expediting computational processes. Further, we add indicators for age at trial start in 5-year brackets from 50 to 75.

Treatments: We consider all ATC level 4 drugs as potential treatments as this level of granularity provides the best trade off in our data between specificity of the drugs used and reasonably sized treatment groups. However, we restrict our analysis to drugs with a minimum of 1000 treated individuals in a given emulated target trial.

Outcomes: Same as for the observational study.

Eligibility criteria: Eligibility criteria are specific for each treatment and outcome. The start date for each trial is the 1st January, 2008. Individuals have to be between the age of 50 and 80 to be able to join. Further, the specific outcome under study should not have occurred prior to the start date. Individuals who have an indication of treatment in the preceding 5 years (1st January, 2003) are excluded.

Treatment assignment: All individuals eligible for the trial on the 1st January, 2008 are assigned a random time, uniformly drawn from 1-36 months. If an individual has an indication of treatment within this time window, the earliest time of treatment initiation is set as the new allocated time for the individual and they enter the treatment arm. If no treatment indication is registered in the time window the individual enters the control arm. Individuals in the treatment arm stay on it during the entire study period. We do not consider treatment discontinuation as it is difficult to define generically valid intervals of treatment intermittence. Individuals in the control arm that subsequently switch to treatment become censored 6 months after the switch. We allow for a small time window to limit possible effects through treatment-by-indication.

Treatment strategies: The strategies to be compared are (i) initiation of treatment and presumed continuation over the study period. (ii) No initiation of treatment during the entire study period.

Follow-up: After treatment determination every individuals goes through a 1 year washout period to avoid the identification of treatment-by-indication effects but also to allow for a phase-in period of the drug. If an event occurred during treatment determination or the washout period, individuals are removed from the study. All remaining individuals are followed for a maximum of 36 months which constitutes the study end or until the occurrence of the event or possible censoring (death, emigration).

Causal estimands: The primary outcome of the study is the average treatment effect between the treatment group and the control group after 36 months, measured as the difference between the absolute risk in the two arms. The causal contrast is the analog of the intention-to-treat.

Statistical analysis: The primary statistical method for the emulated target trial analysis is based on a targeted maximum likelihood estimator (TMLE) for the parameters of longitudinal static and dynamic marginal structural

models as implemented in R-ltmle [22]. For details about TMLE we refer to [21]. One aspect of the current implementation is the need for discretization of the follow-up time, therefore, we split time into 6 months intervals. For each time point 3 effective models are estimate, capturing: (i) the treatment assignment, (ii) the outcome model, (iii) the censoring mechanism, and subsequently combined. Each fit is done via parametric generalized linear models containing the aforementioned covariates plus a treatment indicator where relevant. Otherwise the default parameters for R-ltmle are used.

We extract the TMLE estimate for our primary end point of the 36 months absolute risk between the treatment and control group, plus the corresponding 95% confidence interval. Similarly, we extract the corresponding inverse probability of treatment estimator (IPTW) from the same estimation procedure (automatically estimated for TMLE). Last, we also fitted a simple Cox regression on the emulated target trail data with the covariates and a treatment indicator as an additional comparator.

Contributors

AWJ developed the methods and the study design, conducted the analysis, assembled all figures, and wrote the manuscript. IL and SB provided overall feedback and guidance as well as help with the drug databases. AWJ and LHM conceived the study and accessed and verified the Danish data. LHM and SB supervised the study. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declarations of interests

SB received personal compensation for managing board memberships at Intomics and Proscion and is a scientific advisory board member of Biocenter Finland, Health Data Research UK, the Finnish Center of Excellence in Complex Disease Genetics, ELIXIR Node (Luxembourg), Lund University Diabetes Centre (Lund, Sweden), and SciLifeLab (Stockholm, Sweden). SB reports stocks in Intomics, Hoba Therapeutics Aps, Novo Nordisk, Eli Lilly and Lundbeck. All other authors declare no competing interests.

Data sharing

Danish registry data are available for use in secure, dedicated environments via application to the Danish Patient Safety Authority and the Danish Health Data Authority.
Code is available on https://github.com/alexwjung/DrugTarget

Acknowledgements

This work was supported the Novo Nordisk Foundation under grants NNF17OC0027594 and NNF14CC0001.

References

[1] Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* **11**, 191–200 (2012). URL https://www.nature.com/articles/nrd3681.

[2] Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery* **14**, 475–486 (2015). URL https://www.nature.com/articles/nrd4609.

[3] Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844–853 (2020).

[4] Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **3**, 673–683 (2004). URL https://www.nature.com/articles/nrd1468.

[5] Goldstein, I. *et al.* Oral sildenafil in the treatment of erectile dysfunction. Sildenafil Study Group. *The New England Journal of Medicine* **338**, 1397–1404 (1998).

[6] Singhal, S. *et al.* Antitumor Activity of Thalidomide in Refractory Multiple Myeloma. *New England Journal of Medicine* **341**, 1565–1571 (1999).

[7] Steins, M. B. *et al.* Efficacy and safety of thalidomide in patients with acute myeloid leukemia. *Blood* **99**, 834–839 (2002). URL https://www.sciencedirect.com/science/article/pii/S0006497120382744.

[8] Robins, J. M., Hernán, M. a. & Brumback, B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* **11**, 550–560 (2000). URL https://www.jstor.org/stable/3703997.

[9] Laan, M. J. v. d. & Rubin, D. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics* **2** (2006).

[10] Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* **183**, 758–764 (2016).

[11] Hernán, M. A., Wang, W. & Leaf, D. E. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA* **328**, 2446–2447 (2022).

[12] Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S. & Hernán, M. A. Evaluating Metformin Strategies for Cancer Prevention: A Target Trial Emulation Using Electronic Health Records. *Epidemiology (Cambridge, Mass.)* **34**, 690–699 (2023).

[13] Rein, S. M. *et al.* Integrase strand-transfer inhibitor use and cardiovascular events in adults with HIV: an emulation of target trials in the HIV-CAUSAL Collaboration and the Antiretroviral Therapy Cohort Collaboration. *The lancet. HIV* **10**, e723–e732 (2023).

[14] Szmulewicz, A. G. *et al.* Emulating a Target Trial of Dynamic Treatment Strategies for Major Depressive Disorder Using Data From the STARD Randomized Trial. *Biological Psychiatry* **93**, 1127–1136 (2023). URL https://www.sciencedirect.com/science/article/pii/S0006322322016365.

[15] Hulme, W. J. *et al.* Comparative effectiveness of BNT162b2 versus mRNA-1273 covid-19 vaccine boosting in England: matched cohort study in OpenSAFELY-TPP. *BMJ (Clinical research ed.)* **380**, e072808 (2023).

[16] Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R. & Shrier, I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology* **79**, 70–75 (2016).

[17] Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S. & Hernán, M. A. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine* **25**, 1601–1606 (2019). URL https://www.nature.com/articles/s41591-019-0597-x.

[18] Wallach Kildemoes, H., Toft Sørensen, H. & Hallas, J. The Danish National Prescription Registry. *Scandinavian Journal of Public Health* **39**, 38–41 (2011).

[19] Schmidt, M. *et al.* The Danish National Patient Registry: a review of content, data quality, and research potential. *Clinical Epidemiology* **7**, 449–490 (2015). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655913/.

[20] Jung, A. W. & Gerstung, M. Bayesian Cox regression for large-scale inference with applications to electronic health records. *The Annals of Applied Statistics* **17**, 1064–1085 (2023). URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-17/issue-2/Bayesian-Cox-regression-for-large-scale-inference-with-applications-to/10.1214/22-AOAS1658.full.

[21] Petersen, M. *et al.* Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models. *Journal of Causal Inference* **2**, 147–185 (2014).

[22] Lendle, S. D., Schwab, J., Petersen, M. L. & Laan, M. J. v. d. ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software* **81**, 1–21 (2017). URL https://doi.org/10.18637/jss.v081.i01.

[23] Jung, A. W. *et al.* Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study. *medRxiv* 2022–10 (2022).

[24] Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *The Lancet* **344**, 1383–1389 (1994). URL https://www.sciencedirect.com/science/article/pii/S0140673694905665.

[25] Ridker, P. M. *et al.* Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. *New England Journal of Medicine* **359**, 2195–2207 (2008). URL https://doi.org/10.1056/NEJMoa0807646.

[26] Olmastroni, E. *et al.* Statin use and risk of dementia or Alzheimer's disease: a systematic review and meta-analysis of observational studies. *European Journal of Preventive Cardiology* **29**, 804–814 (2022). URL https://doi.org/10.1093/eurjpc/zwab208.

[27] Petek, B. *et al.* Statins and cognitive decline in patients with Alzheimer's and mixed dementia: a longitudinal registry-based cohort study. *Alzheimer's Research & Therapy* **15**, 220 (2023). URL https://doi.org/10.1186/s13195-023-01360-0.

[28] Ren, Q.-w. *et al.* Statins and risks of dementia among patients with heart failure: a population-based retrospective cohort study in Hong Kong. *The Lancet Regional Health – Western Pacific* **44** (2024). URL https://www.thelancet.com/journals/lanwpc/article/PIIS2666-6065(23)00324-3/fulltext.

[29] Mundal, L. J. *et al.* Association of Familial Hypercholesterolemia and Statin Use With Risk of Dementia in Norway. *JAMA Network Open* **5**, e227715 (2022). URL https://doi.org/10.1001/jamanetworkopen.2022.7715.

[30] Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Nature Scientific Data* (2023). URL https://www.nature.com/articles/s41597-023-01960-3.

[31] Schrezenmeier, E. & Dörner, T. Mechanisms of action of hydroxychloroquine and chloroquine: implications for rheumatology. *Nature Reviews Rheumatology* **16**, 155–166 (2020).

[32] Varma, V. R. *et al.* Hydroxychloroquine lowers alzheimer's disease and related dementias risk and rescues molecular phenotypes related to alzheimer's disease. *Molecular psychiatry* **28**, 1312–1326 (2023).

[33] Skwarski, M. *et al.* Mitochondrial inhibitor atovaquone increases tumor oxygenation and inhibits hypoxic gene expression in patients with non–small cell lung cancer. *Clinical Cancer Research* **27**, 2459–2469 (2021).