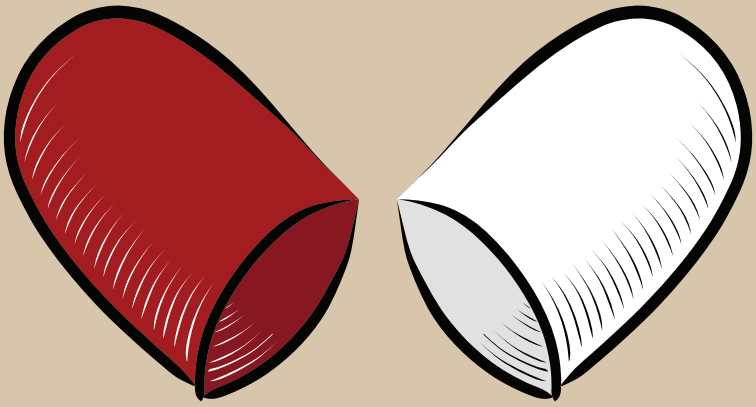
UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

THESIS



PHARMACOVIGILANT MACHINE LEARNING IN BIG DATA?

BENJAMIN SKOV KAAS-HANSEN MD, MSc

THIS THESIS HAS BEEN SUBMITTED TO
THE GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES, UNIVERSITY OF COPENHAGEN
4 NOVEMBER 2021

PHARMACOVIGILANT MACHINE LEARNING IN BIG DATA?

The work presented in this thesis was carried out between 1 October 2017 and 31 October 2021 at the Clinical Pharmacology Unit (Zealand University Hospital, Denmark) and the Novo Nordisk Foundation Center for Protein Research (University of Copenhagen, Denmark). The PhD project was funded by the Novo Nordisk Foundation (NNF17OC0027594, NNF14CC0001) and Innovation Fund Denmark (5153-00002B).

CANDIDATE

Benjamin Skov Kaas-Hansen MD, MSc
Clinical Pharmacology Unit, Zealand University Hospital
Novo Nordisk Foundation Center for Protein Research, University of Copenhagen

SUPERVISORS

Stig Ejdrup Andersen MD, PhD (principal supervisor)
Associate professor, Clinical Pharmacology Unit, Zealand University Hospital

Søren Brunak MSc, PhD (principal co-supervisor)
Professor, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen

Gesche Jürgens MD, PhD
Associate professor, Clinical Pharmacology Unit, Zealand University Hospital

ASSESSMENT COMMITTEE

Morten Andersen MD, PhD (chair)
Professor, Pharmacovigilance Research Center, University of Copenhagen

Erzsébet Horváth-Puhó MSc, PhD
Associate professor, Department of Clinical Epidemiology, Aarhus University

Marco Tuccori PharmD, PhD
Drug safety manager, Unit of Adverse Drug Reactions Monitoring, University Hospital of Pisa
External docent, Department of Pharmacy & Department of Clinical and Experimental Medicine, University of Pisa

Layout, typesetting and illustrations (unless stated otherwise) by Benjamin Skov Kaas-Hansen. Cover image by Simona Gentile.

ISBN (printed version): 978-87-93510-83-8. DOI (electronic version): 10.5281/zenodo.6284793.

THESIS

PHARMACOVIGILANT MACHINE LEARNING IN BIG DATA?

BENJAMIN SKOV KAAS-HANSEN

GRADUATE SCHOOL OF HEALTH SCIENCES
UNIVERSITY OF COPENHAGEN

Contents

List of Figures

Scope and strategy

In 1997 drug safety stakeholders from 34 countries released the Erice Declaration stating that "monitoring, evaluating and communicating drug safety is a public-health activity with profound implications" [1, annex 4]. For several reasons, post-marketing drug safety surveillance is becoming increasingly important; for example accelerating innovation and pressure to let drugs enter the market before efficacy is fully established [2, 3] result in less knowledge about drugs at market entry.

Medicines are used by many: about one in eight Danes (and more than half of the elderly) uses 5 drugs or more concurrently [4, fig. 2]. With an ageing population, increasing use of medicines and accelerated drug-development, strong pharmacovigilant systems to safeguard patients are growing ever more important.

Indeed, there is a downward trend in the time from market entry to first adverse drug reaction (ADR) report, in drugs eventually withdrawn [5]:

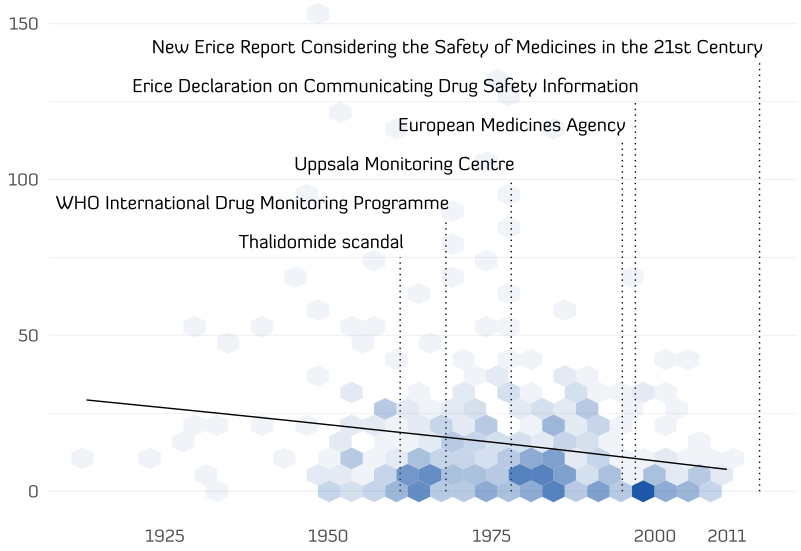


Figure 1: The year of first adverse drug reaction report (x axis) and time on the market (y axis), for drugs that were eventually withdrawn. Observations binned in hexagons to avoid overplotting; the darker the colour the more observations. The non-linear black trend curve uses shrinkage cubic splines [6]. Data from Onakpoya et al. [5], layout inspired by Rodríguez [7].

Albeit an extreme case, the experience with fast development of curative treatments and vaccines during the COVID-19 pandemic emphasised the importance of drug safety surveillance systems to keep up with the development of new drugs and vaccines, and to ensure safe use of existing drugs when indications change [8, 9].

Danish data are widely known and recognised for their high quality, simple (and certain) record linkage, and coverage of many domains. Therefore, examples of pharmacovigilance studies using these registers abound. Large-scale combinations of Danish register data with in-patient data from electronic medical records (EMRs), however, are far less common and the pharmacovigilant potential therein was somewhat unexplored despite the New Erice Report's calling for observational research and emphasising i.a. that pharmacovigilance research serve the interest of the public [10].

Scope

This project was part of the BigTempHealth research programme that collated free text, biochemical and medication data from EMRs in an operational format and linked these with longitudinal data from nation-wide registers going back decades. As such, with this project we sought to understand if and how we can harness massive, secondary observational clinical and administrative data for pharmacovigilance, taking point of departure in idiosyncratic strengths and weaknesses of our infrastructure and setup. Specifically, we sought qualified answers to the following questions:

- Can observational data cater for pharmacovigilance?
- What are the hindrances?
- Is it worth our while?

Strategy

Due to the immensity of this scope, we chose three use cases for answering substantial pharmacovigilance questions with machine learning methods. Each use case became a full study, reported in its own scientific article. We chose different methodologies to cover a substantial part of the methodological spectrum of pharmacovigilance.

Structure

The main content is divided into three parts with three chapters each. Part I provides the necessary background to read and appraise the rest of the thesis and the manuscripts. I first introduce concepts central to the scope of this project and, then, our principal data sources. This part ends with details on core methods used in the three studies when such information did not fit in the manuscripts.

Part II begins with an outline of key methods and main results in studies at a glance. Then, the discussion addresses select overarching challenges and limitations and, to put these into a broader

context of pharmacovigilance. I wrap up this part in concluding remarks and outlook with answers to the questions posed above and pointers to future endeavours, some of which I would like pursue myself.

Part III has one chapter per study. Prevalence and adverse outcomes of drug-drug interactions presents descriptive statistics on the landscape of drug-drug interactions (DDIs) and our attempts at investigating which be associated with adverse outcomes. Renal dysfunction and risk of inappropriate drug dosing reports on the development, and internal validation, of prediction models to identify patients with renal dysfunction at elevated risk of receiving inappropriate doses of select renal risk drugs. Finally, Language-agnostic signal detection in clinical notes presents a novel take on mining clinical narratives for single-drug and DDI safety signals.

In addition to this one, two chapters come before the main content. Summaries contains the scientific abstract and a popularised summary in Danish. Publications and manuscripts lists scientific outputs of mine and as co-author in the course of this PhD project.

Summaries

Abstract

A key component of pharmacovigilance is to ascertain that the medicines we use are safe, and discover if they turn out not to be. In the past decades, there has been growing interest in leveraging new data sources for pharmacovigilance, and especially secondary observational data have received much attention.

Scope

We undertook this project, as part of the BigTempHealth research programme, to understand if and how we can harness massive, heterogeneous patient and register data for pharmacovigilance. We formalised this endeavour with three questions: can observational data cater for pharmacovigilance? What are the hindrances? Is it worth our while?

To answer these questions we employed three substantial pharmacovigilance use cases leveraging data from electronic medical records and nationwise registers, combining methods from (pharmaco)epidemiology, machine learning, natural language processing and data science.

Methods and results

The goal of study I was two-fold. We first described the landscape of drug-drug interactions in a large Danish cohort of 2.9 million admissions of 945,000 patients. Then, matching on high-dimensional preference scores, we used Poisson and Cox regression models to estimate the effect of exposure to discouraged drug pairs on length-of-stay, mortality and readmission. We found that well-known potential drug-drug interactions still abound, and our results suggest, in particular, that prescribing clinicians be alert when using strong inhibitor/inducer drugs (i.e. clarithromycin, valproic acid, terbinafine) and prevalent anticoagulants (i.e. warfarin and NSAIDs) due to their great potential for harmful interactions. Our finding that 3A4 was the most prominent cytochrom P450 isoenzyme involved in mortality and readmission rates agrees well with empirical evidence and clinical experience.

In study II we developed, and internally validated, 10 prediction models in a cohort of 52,451 admissions using multilayer percep-

trons to flag patients with renal dysfunction at elevated risk of various extents of inappropriate dosing of select renal risk drugs. The study leveraged in-patient data with high temporal granularity on drug dispensations and clinical biochemistry to operationalise the outcome variable. We found that the trained prediction models can flag patients at high risk of receiving at least one inappropriate dose daily. The multilayer perceptrons performed slightly better than their ridge regression counterparts with respect to calibration while discrimination was similar for the models. Using a method from the field of explainable artificial intelligence we confirmed that multilayer perceptron models picked up non-linear effects that escaped the ridge regression models.

With study III we propose a novel, complementary method for safety signal mining in structured medication data and clinical free text, regardless of language and with no need for prior manual curation. This end-to-end pipeline is based on word embeddings and uses multilayer perceptrons as the core for association elicitation. We used data from 2.9 million inpatient visits featuring 10.8 million clinical notes and 13.7 million drug prescriptions. In total, we trained 10,720 multilayer perceptron models using almost 180 million tokens retained from the clinical notes. After manual review of the safety signals, based on well-established reference sets, we found the method's *hit rate* was in the same order of magnitude as that in individual case safety reports, the longstanding mainstay of safety signal detection.

Conclusion

We conjencture that secondary observational data—in particular, combined data from patient records and from national registers—are of genuine utility in pharmacovigilance, but only if several hindrances are overcome or at least accounted for upfront: the data can be misleading (due to erroneous and/or missing data for various reasons), disparate (requiring standardisation and integration), and complex (necessiting particular skillsets and organisational infrastructures to soundly leverage the data).

Putting these kinds of data on a form appropriate for pharmacovigilance is worthwhile, but it should be done in a way that facilitates scrutiny and transparency, reduces the risk of errors, prevents duplicative work, and enables international collaboration. To this end, Danish observational data should be readily available in appropriate common data models.

Further, our studies and methodological advances in general suggest that machine-learning driven safety signal detection, causal inference and causal discovery are likely viable complements to safety signal detection in individual case safety reports.

Populærvidenskabeligt sammendrag

Et centralt mål med pharmacovigilance—den videnskabelige disciplin, der beskæftiger sig med medicinbivirkninger—er at forvisse sig om, at medicin er sikker at bruge og opdage det, hvis det ikke er tilfældet. Traditionelt bygger dette på såkaldte spontane bivirkningsindberetninger fra sundhedspersonale, patienter og medicinalvirksomheder, men i de senere år er man begyndt at søge efter alternative og supplerende datakilder. Her har interessen især været rettet mod såkaldte sekundære observationelle data, der opsamles som biprodukt af eksempelvis journalregistreringer under hospitalsindlæggelser og ikke særligt i forskningsøjemed.

Projektets formål

Dette projekt var del af forskningsprogrammet BigTempHealth. Formålet var at undersøge om og forstå hvordan store, forskelligartede datasæt fra patientjournaler og nationale registre kan udnyttes i pharmacovigilance. For at indsnævre dette formulerede vi tre spørgsmål: Kan observationelle data bruges i pharmacovigilance? Hvilke forhindringer er der? Kan det betale sig?

Til at besvare disse spørgsmål designede vi tre videnskabelige studier, der gjorde brug af data fra elektroniske patientjournaler og nationale registre samt analytiske metoder fra (farmako)epidemiologi, machine learning, natural language processing (statistiske metoder, der gør os i stand til at bruge tekstdata i kvantitative analyser) og datavidenskab.

Metoder og resultater

Målet med studie I var todelt. Først kortlagde vi omfanget af medicininteraktioner i en stor dansk kohorte med 2.9 millioner indlæggelser af 945,000 patienter. Dernæst undersøgte vi, om eksponering for medicinkombinationer, der bør undgås, var associeret med henholdvis øget dødelighed, øget risiko for genindlæggelse og forlænget indlæggelse. Vi fandt, at præparater med velkendte interaktioner stadig gives sammen i relativt vid udstrækning, og vore resultater tyder på, at læger særligt bør være opmærksomme ved brug af stærke hæmmere/inducere (såsom clarithromycin, valproat og terbinafin) og almindeligt brugte blodfortyndende stoffer (warfarin og NSAIDs) på grund af det store potentiale for skadelige bivirkninger. Vi fandt også, at 3A4 var det mest prominente såkaldte CYP-isoenzym (involeret i nedbrydning af medicin i leveren) i relation til øget mortalitet og risiko for genindlæggelse, hvilket stemmer fint overens med den videnskabelige litteratur og kliniske erfaring.

I studie II brugte vi en kohorte på 52.451 indlæggelser til at udvikle og validere 10 prædiktionsmodeller, på basis af klassiske metoder og neurale netværk, til at identificere patienter med nyresvigt i øget risiko for at modtage uhensigtsmæssige doser

af såkaldte *renal risk drugs*, altså stoffer kendt for at være problematiske i forbindelse med netop nyresvigt. I studiet udnyttede vi den høje tidsopløsning i indlæggelsesdata for medicinadministrationer og målinger af nyrefunktionen. Vi fandt, at modellerne kunne identificere patienter med forhøjet risiko for at modtage mindst 1 uhensigtmæssig dosis dagligt. Vore neurale netværksmodeller var lidt bedre end de klassiske, og vi kunne vise, at førstnævnte var i stand til at opfange mere komplekse associationer i vore data end sidstnævnte.

Med studie III foreslår vi en ny, supplerende metode til at detektere såkaldte sikkerhedssignaler i strukturerede medicindata og naturlig tekst fra journalnotater, uafhængigt af sprog og uden behov for forudgående manuel kodning af teksten. Metoden baserer sig på en såkaldt embeddingmodel, der kan konvertere ord og sætninger til numeriske data, så eksempelvis ord med samme eller næsten samme betydning kommer til at "ligne hinanden" uanset den oprindelige stavemåde (f.eks. vil *hovedpine* og *hovedsmerter* ligne hinanden men være forskellige fra *hoste*). Vi brugte data fra 2.9 millioner indlæggelser: 13.7 millioner medicinordinationer og næsten 180 millioner tekstbidder bibeholdt fra 10.8 millioner journalnotater efter oprensning og filtrering. I alt brugte vi 10.720 neurale netværksmodeller til at detektere sikkerhedssignaler. Vor manuelle kontrol af de vigtigste signaler bekræftede en *hit rate* i samme størrelsesorden som man ser i spontane bivirkningsindberetninger, der traditionelt har udgjort og stadig udgør den primære kilde til sikkerhedssignaler.

Konklusion

Vi fandt således holdepunkter for, at sekundære observationelle data—særligt kombinationen af data fra patientjournaler og nationale registre—kan bruges i pharmacovigilance. Det kræver dog, at adskillige forhindringer overvindes, eller at der fra starten på anden måde tages højde for dem: Sådanne data kan være misvisende (eksempelvis som følge af fejlagtige og/eller manglende data), adskilte (og derfor skal integreres og standardiseres) og komplekse (hvilket blandt andet kræver en organisation med forskelligartede kompetencer og adgang til adækvat hardware).

Vi mener altså, det kan betale sig at gøre brug af sekundære observationelle data i pharmacovigilance, men det bør ske på en måde, der faciliterer granskning og gennemsigtighed, sænker risikoen for fejl, forebygger dobbeltarbejde og giver mulighed for internationale samarbejder om at lære så meget som muligt fra de data, vi har adgang til. For at sikre dette, bør danske data være tilgængelige i standardformater, som ikke kun bruges i Danmark.

Slutteligt tyder resultaterne af vore studier og generelle metodologiske fremskridt på, at machine learning-drevet detektion af sikkerhedssignaler, kausalinferens og causal discovery kan udgøre et reelt supplement til spontane bivirkningsindberetninger.

Publications and manuscripts

Study I

Rodríguez CL[§], **Kaas-Hansen BS**[§], Eriksson R, Hernansanz JB, Belling KG, Andersen SE, Brunak S. *Drug interactions in hospital prescriptions in Denmark: Prevalence and associations with adverse outcomes. In review.*
§ shared first-authorship
medRxiv pre-print available (10.1101/2021.05.27.21257764)
An earlier version of this manuscript was used in Ms. Rodríguez' PhD thesis [7]

Study II

Kaas-Hansen BS, Rodríguez CL, Placido D, Thorsen-Meyer H-C, Nielsen AP, Dérian N, Brunak S, Andersen SE. *Identifying patients at high risk of inappropriate drug dosing in periods with renal dysfunction. In review.*
medRxiv pre-print available (10.1101/2021.07.09.21257018)

Study III

Kaas-Hansen BS, Placido D, Rodríguez CL, Gentile S, Thorsen-Meyer H-C, Nielsen AP, Brunak S, Jürgens G, Andersen SE. *Eliciting side effects from clinical notes: language-agnostic pharmacovigilant text mining. In preparation.*

Not included: published or accepted

1. Granholm A, **Kaas-Hansen BS**, Kjær M-BN, ..., Hyllander M (2021). *Patient-important outcomes other than mortality in recent ICU trials: protocol for a scoping review*. Acta Anaesthesiologica Scandinavica (10.1111/aas.13937).
2. Reps JM, Kim C, Williams RD, ..., **Kaas-Hansen BS**, ..., Rijnbeek PR; for the OHDSI-COVID19 consortium (2021). *Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study*. JMIR Medical Informatics (10.2196/21547).
3. Jimenez-Solem E, Petersen TS, Hansen C, ..., **Kaas-Hansen BS**, ..., Sillesen M (2021). *Developing and validating a COVID-19 adverse outcome risk prediction model from a bi-national European cohort of 5594 patients*. Scientific Reports (10.1038/s41598-021-81844-x).
4. Jürgens G, Andersen SE, Rasmussen HB, Werge T, Jensen H, **Kaas-Hansen BS**, Nordentoft M (2020). *Effect of routine CYP450 2D6 and 2C19 Genotyping on Antipsychotic Drug Persistence in Patients with Schizophrenia – an investigator initiated randomized controlled utility trial*. JAMA Network Open (10.1001/jamanetworkopen.2020.27909).

5. Lane JCE, Weaver J, Kostka K, ..., **Kaas-Hansen BS**, ..., Prieto-Alhambra D; for the OHDSI-COVID-19 consortium (2020). *Risk of depression, suicidal ideation, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multi-national network cohort study*. Rheumatology (10.1093/rheumatology/keaa771).
6. Burn E, You SC, Sena A, ..., **Kaas-Hansen BS**, ..., Ryan P; for the OHDSI-COVID-19 consortium (2020). *Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study*. Nature Communications (10.1038/s41467-020-18849-z).
7. Lane JCE, Weaver J, Kostka K, ..., **Kaas-Hansen BS**, ..., Prieto-Alhambra D; for the OHDSI-COVID-19 consortium (2020). *Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study*. Lancet Rheumatology (10.1016/S2665-9913(20)30276-9).
8. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, **Kaas-Hansen BS**, Toft P, Schierbeck J, ..., Perner A (2020). *Dynamic and explainable machine learning prediction of mortality in ICU patients: a retrospective study of high-frequency data in electronic patient records*. Lancet Digital Health (10.1016/S2589-7500(20)30018-2).
9. Granholm A, Marker S, Krag M, Zampieri FG, Thorsen-Meyer HC, **Kaas-Hansen BS**, ..., Perner A (2020). *Heterogeneity of treatment effect of prophylactic pantoprazole in adult ICU patients: a post hoc analysis of the SUP-ICU trial*. Intensive Care Medicine (10.1007/s00134-019-05903-8).
10. Graudal N, **Kaas-Hansen BS**, Guski L, Hubeck-Graudal, Welton N, Jürgens G (2019). *Different original and biosimilar TNF inhibitors similarly reduce joint destruction in rheumatoid arthritis. A network meta-analysis of 36 randomized controlled trials*. International Journal of Molecular Sciences (10.3390/ijms20184350).
11. Granholm A, Marker S, Krag M, Zampieri FG, Thorsen-Meyer HC, **Kaas-Hansen BS**, ..., Perner A (2019). *Heterogeneity of treatment effect of stress ulcer prophylaxis: protocol for an exploratory secondary analysis*. Acta Anaesthesiologica Scandinavica (10.1111/aas.13432).

Not included: preprints or in preparation

1. Thorsen-Meyer H-C, Placido D, **Kaas-Hansen BS**, ..., Brunak S. *Discrete-time survival analysis in the critically ill: an interpretable deep learning approach using heterogeneous data from electronic patient records and national registers*. (In preparation).
2. Placido D, ..., **Kaas-Hansen BS**, ..., Brunak S. *Dynamic prediction of risk of unplanned ICU admission and mortality in hospitalized patients*. (In preparation, tentative order of co-authors).
3. Nielsen AP, Placido D, **Kaas-Hansen BS**, ..., Brunak S. *Predicting individual risk of intrapartum caesarean delivery after the start of the first and second phase of labour using explainable machine learning: a development and validation study using registry data and clinical records*. (In preparation).
4. Granholm A, ..., **Kaas-Hansen BS**, ..., Hylander-Møller M. *Health-related quality of life and days alive without life support or out of hospital: protocol*. (In preparation, tentative order of co-authors).
5. Granholm A, ..., **Kaas-Hansen BS**, ..., Hylander-Møller M. *Modelling "days alive without"-type outcomes in randomised clinical trials*. (In preparation, tentative order of co-authors).
6. Granholm A, **Kaas-Hansen BS**, ..., Hylander-Møller M. *Patient-important outcomes other than mortality in recent ICU trials: scoping review*. (In preparation).
7. **Kaas-Hansen BS**, Graudal N, Andersen SE, Thygesen LC, Brunak S, Jürgens G (2020). *Hyponatraemia and mortality in psychiatric patients: protocol for Bayesian causal inference study*. medRxiv (not peer-reviewed, 10.1101/2020.06.25.20138206).

Other outputs: software and analytic code

1. **Kaas-Hansen BS**. NetworkMetaAnalysis: A package in the OHDSI ecosystem to take a collection of target-comparator-outcome estimates and produce indirect network meta-analytic estimates across all target-comparator-outcomes. Software (R-package). Available from github.com/OHDSI/NetworkMetaAnalysis (under active development in the develop branch).
2. **Kaas-Hansen BS**, Thomsen AH. humapr: R package for visualising topographic human data. Software (R-package). Available from github.com/epiben/humapr.
3. **Kaas-Hansen BS**. promedreadr: A utility R package for extracting select information from pro.medicin.dk. Software (R-package). Available from github.com/epiben/promedreadr.
4. **Kaas-Hansen BS**, Rodríguez CL, Placido D. Analytic code for study II. Analytic code (R, Python, SQL). doi:10.5281/zenodo.4560078
5. **Kaas-Hansen BS**, Placido D, Rodríguez CL. Analytic code for study III. Analytic code (R, Python, SQL). doi:10.5281/zenodo.5598068
6. **Kaas-Hansen BS**, Derian N, Drakos I, Ostropolets A, Davydov A, Andersen SE. An OMOP-based tool for surveying and visualising concurrent drug exposure and renal function. Software (Shiny app). doi:10.5281/zenodo.5609153. *Presented at the 2nd European OHDSI Symposium 2019* [11]

I

Concepts

In this first chapter, I define key concepts and explain terminology used throughout. Some concepts are left out because sufficient details are available in the manuscript(s).

Chapter contents

Adverse drug events and reactions

The common term *side effect* has two more technical cousins with specific definitions: adverse drug event (ADE) and adverse drug reaction (ADR). ADEs comprise any noxious event following exposure to a medicine with their temporal order as the only qualifying characteristic. ADRs are reactions to drug exposure as a result of known or plausible biological causal pathways [12, 13] and, as such, constitute a subset of ADEs.

ADRs are usually classified in 6 groups [14] (ABCDEF, mnemonic in parentheses):

- Dose-related (augmented)
- Not dose-related (bizarre)
- Dose-related and time-related (chronic)
- Time-related (delayed)
- Withdrawal (end of use)

- Unexpected failure of therapy (failure)

Dose-related reactions, for example, are somewhat predictable because we (mostly) know the pharmacodynamic profile of the drug. This kind of ADR is among the targets of safety studies as part of the drug development pipeline [15]. In contrast, not dose-related ADRs are more unpredictable and tend to not be related to the pharmacological effect of the drug. This makes them somewhat more interesting from a pharmacovigilance and safety signal detection point of view because their identification necessitates systems able to detect unexpected relationships.

Pharmacovigilance

The lack of systematic surveillance of exposures and noxious outcomes allowed the thalidomide[1] scandal to fly under the radar for years and the disaster to unfold [16]. Although pharmacovigilance really took off at an international scale in the wake of, and as a response to the thalidomide scandal, Sweden and the United Kingdom (UK), for example, already had drug safety systems in place [17]. The concerted effort now widely in place came about especially with the Europeanisation of drug regulation [17] and Uppsala Monitoring Centre (UMC)[2] has played a key role in building capacity in national agencies across the globe to maintain drug reporting systems [1].

[1] Incidentally, thalidomide was among the safety signals undergoing manual review in study III

[2] The World Health Organization Collaborating Centre for International Drug Monitoring, established in 1978

Indeed, since its inception pharmacovigilance has grown in scope to become "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" [18] with 4 objectives [19]:

- Improve patient care and safety in relation to the use of medicines, and all medical and paramedical interventions.
- Improve public health and safety in relation to the use of medicines.
- Contribute to the assessment of benefit, harm, effectiveness and risk of medicines, encouraging their safe, rational and more effective (including cost-effective) use.
- Promote understanding, education and clinical training in pharmacovigilance and its effective communication to health professionals and the public.

Spontaneous reporting system

Systematic collection of individual case safety reports (ICSRs) still contitutes the mainstay of post-marketing drug safety surveillance in modern pharmacovigilance [16, 17, 20]; for exampe, as of November 2019, VigiBase (maintained by UMC) marshalled ICSRs from more than 130 countries [21]. EudraVigilance, maintained

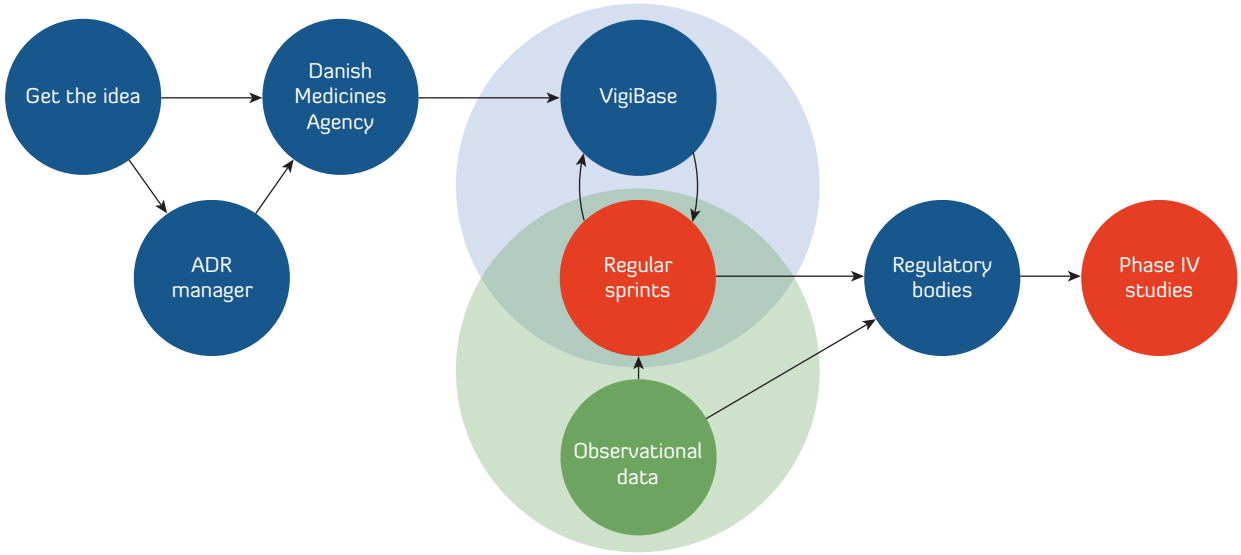


Figure 2: Schematic illustration of the flow of spontaneous case reporting system, from clinical suspicion until the ICSR lands in a database (VigiBase in this example).

by the European Medicines Agency (EMA), collects ICSRs from relevant authorities in European Union (EU) member states as well as marketing authorisation holders [22]; the FDA Adverse Event Reporting System (FAERS) has a similar mandate in the United States of America (USA) [23]. Because this system was already in place, it served the public well amid the COVID-19 pandemic, helping elicit problems with i.a. hydroxychloroquine [24] and the Vaxzevria COVID-19 vaccine [25].

Figure 2 depicts the flow of information in spontaneous reporting systems (SRSs). Submitting an ICSR is mandatory in certain cases. In Denmark, for example, reporting is mandatory if the suspected ADR occurs within two years of market authorisation or when the medicine is subject to so-called *stricter reporting requirements* (e.g. benzodiazepine and opioids); the Danish Medicines Agency (DMA) maintains a list of medicines that fall in the latter group[3] [28].

It is a key strength of ICSRs that they only exist because someone suspected an ADR, but they suffer from several weaknesses. Two such are their potentially limited longitudinal information and lack of data on concurrent exposure to other drugs, which may impede causality evaluation and hamper safety signal detection. Thus, it is often impractical to exhaustively study exposure over time because not all submitted ICSRs will provide full, or even sufficient accounts of drug exposure trajectories.

Another weakness is the high degree of underreporting, with one—albeit a bit old—estimate of 94% [30]. More recent data from DMA (figure 3) show stably few ICSRs in Denmark. The reasons for underrepoting has been studied intensely over the years, but seem difficult to remedy. In Denmark, for example, the departments of clinical pharmacology have established regional ADRs managers, a mechanism to lower the barrier of reporting suspicions of ADRs

[3] Incidentally, despite its opioid nature, tramadol was not subject to stricter reporting requirements for a long time as it was believed be less addictive, a pharmacologically questionable claim, being a pro-drug with the active agent binding to μ-receptors [26, accession number: DB00193] and given in morphine-equianalgetic doses [27]

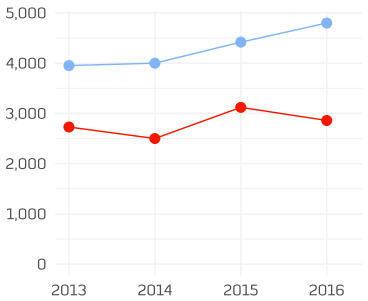


Figure 3: The number of ICSRs to DMA (y axis) by year (x axis). Red: serious reports. Blue: not serious reports. Data from [29, fig. 1].

[31–33].

Safety signals

Several definitions of safety signals exist [34–36], but I have used the original definition of the World Health Organization (WHO) even if is unclear whether a safety signal arising from observing thousands of patients constitutes only a single signal:

> "Reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously. Usually more than a single report is required to generate a signal, depending on the seriousness of the event and the quality of the information." [37]

Safety signals generally hinge on four data sources: ICSRs [20, 38], online forums (including social media) [39–52], scientific literature [20] and longitudinal patient data [38, 53–57]. The analytical approaches are characterised by varying levels of modelling complexity and data structuredness.

Electronic medical records and electronic health records

Although often used interchangeably, electronic medical records (EMRs) and electronic health records (EHRs) are distinct concepts [58, 59]. Considering electronic patient records (EPRs) and EMRs synonyms and leaning on a common definition[4] [60, 61], I use the following definitions:

- EMRs: the data recorded during hospital visits[5] (such as drug administrations and biochemical values measured) and stored in a system underpinning clinical care. These records also serve a legal purpose in that they document when what was observed and done, and by whom.
- EHRs: the data recorded in e.g. the national registers, so a collection of (usually structured) data from various sources in a unified (if possibly idiosyncratic) data model. EMRs serve as a key data source of EHR data.

[4] Although many cite Habib [60] (see e.g. Google Scholar), the article now seems available only through Wayback Machine

[5] Could also be e.g. at a family doctor, but we only had in-hospital data

Structured and unstructured data

Clinical data come about by highly structured data collection processes. Many patient data are stored in a structured format, for example, drug prescriptions (with e.g. timestamps for start and end of treatment, dosing, and administration instructions) and biochemistry (e.g. timestamp, specimen, and result) So by *structured data* I mean information stored in a tabular format, be it wide or long.

Defining *unstructured data* is more challenging and might most easily be defined as its lack of structure. In the domain of patient

data, clinical notes are the most prominent unstructured data. Thus, depending on the task at hand, different natural language processing (NLP) methodologies must be invoked to make free text play well with methods for quantitative analyses.

Drug-drug interactions

Drug-drug interactions (DDIs) occur if one drug changes the action of another when used together [62, ch. 47] and fall in two categories: pharmacokinetic and pharmacodynamic. The number of potential DDIs grows fast with the number of drugs used concurrently (figure 4) and DDIs may be involved in 10%-30% of ADRs [62, 63]. DDIs constitute a key facet of pharmacovigilance: they constitute a minor group of ADRs but are to some extent predictable (making them somewhat preventable or at least manageable) [63], and their bidimensional nature poses several challanges to modelling, as discussed in Outcome modelling and Outcome operationalisation.

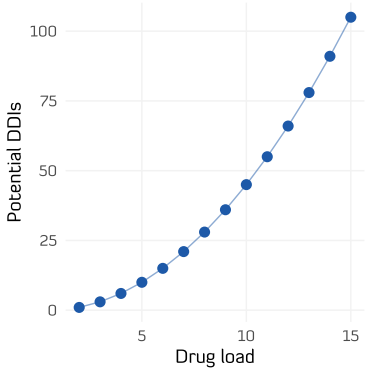


Figure 4: The number of potential drug-drug interactions (y axis) increases fast with the drug load (= number of drugs used concurrently, x axis).

Epidemiological enquiry

In the area of machine learning, confusion seems to exist about epidemiological methods and study designs, seemingly because statistical models can be (and are) used to answer different questions depending on the study design [64]. Thus, with this section and the next (Statistical modelling) I try to disentangle some common misconceptions in this regard.

The common contrasting of epidemiology with machine learning is misguided [65, 66]: epidemiological enquiry is defined by its scope, not its methods, and as such machine learning is one type of tools to underpin such enquiry. Indeed, epidemiology can be defined as "the study of the occurrence and distribution of health-related states or events in specified populations, including the study of the determinants influencing such states, and the application of this knowledge to control the health problems" [67] and falls in three categories: descriptive, analytical and interventional [64, 68].

Controlled trials mimic the laboratory in which the scientist has full control (apart from aleatoric uncertainty) over the setting including which subjects are exposed and which are not. This way, they can elicit truly causal effects of the exposure by (ideally) keeping all other factors identical or at least identically distributed. The randomised controlled trial (RCT) is perhaps the most widely used because we do not have access to "knock-out" humans who are identical except for the exposure, and so randomisation ensures that all factors except the exposure be distributed equally: with a sufficiently large study population, the difference in outcome between exposed and non-exposed subjects is attributable to the exposure.

Descriptive epidemiology makes no attempt at eliciting causal factors of exposures but solely describes trends, patterns or (co)occurrences

of phenomena. Descriptive studies aid hypothesis generation and can hint at interesting associations but temporality must be considered to gauge the strength and relevance of associations in observational studies.

Analytical epidemiology, essentially, comes in two forms: aetiological and predictive. Aetiological studies seek to answer the same questions as RCTs but in settings where the scientists holds no sway over exposure status, i.e. they must resort to observational data. This can be used when planning an RCT to obtain realistic effect-size estimate to inform study planning or when there is no ethically, financially or logistically defensible way to conduct an RCT to gauge the effect size.

Statistical modelling

"All models are wrong but some are useful"[6] and, conceptually, we can think of statistical models in a modular fashion with three components. The first component consists of data transformations to put data in an operational format, most often by submitting the raw data to (hefty) preprocessing. This is usually called feature engineering in machine learning contexts. I use the term *feature* to refer to the transformed data variable appropriate as input in machine learning models. Sometimes, data transformation is subsumed into the full model, for example when using embeddings [69–71].

[6] Often attributed to George E. P. Box

The second component models the data-generating process in some way. In aetiological epidemiology, when we build models to elicit effect-size estimates (risk ratios, odds ratio, hazard ratios, etc.), this can be called explanatory modelling [72]. In contrast, predictive epidemiology uses models that seek not to to mimic the world but, rather, to approximate it sufficiently well to make predictions as close to the observations as possible [72].

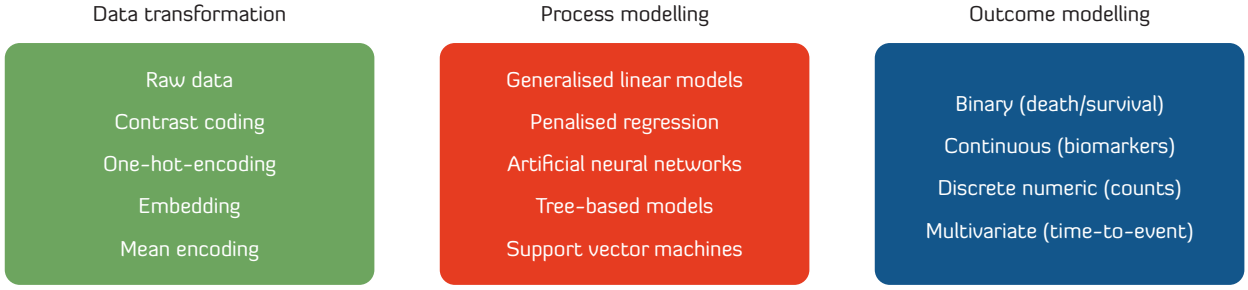


Figure 5: Conceptual, modular approach to statistical modelling. Examples of outcomes in parentheses. Inspired by i.a. Shmueli [72].

Most applications of machine learning in epidemiology are purely predictive. In recent years researchers increasingly seek to explain how results of prediction models come about, often referred to as *opening the black box*, with technologies from explainable artificial intelligence (xAI) such as SHapley Additive exPlanation (SHAP) values [73]. Even the best explanations, however, will not

render the models explanatory and alternative methodologies, e.g causal discovery, are required to bridge this gap.

The final component is the outcome. A classic logistic regression, for example, models a binary outcome on the log-odds scale instead of the probability scale directly (see page 40) while modelling discrete, non-negative numeric outcomes (e.g. length-of-stay) requires the model to handle this restriction correctly. Outcome modelling also entails deciding whether to use e.g. maximum likelihood estimation [74] or other loss functions (often, nonetheless, based on the likelihood function [75, ch. 6]). The term *label* means the observed value in the outcome variable for a given unit, for example that patient A deceased in the follow-up peirod.

Machine learning

Statistics and machine learning have similar objectives but owing to their different origins (mathematics and computer science), equivalent or similar concepts have taken on different names [74]. The term *machine learning* came about in 1959, in a paper on programming a computer program to play checkers better than its creator [76], and usually divides problems into unsupervised and supervised [77].

Machine learning models, and no less neural network models, are trained by minimising the loss [70]. The loss function determines entirely what the model will learn and what it will focus on while learning. These models are *ruthless learners* so defining the right loss function is crucial: unless explicitly specificed not to, for example, mistaking one benign tumor from another will be considered no worse than mistaking a malignant tumor from a benign [78]. One immediate consequence is that machine learning models will learn, and thus perpetuate or even corroborate, biases captured in the input data [79, 80].

The term *ruthless learners* due to Stuart Russell [78]

Unsupervised learning

Unsupervised problems have no ground truth: there are no labels to predict, making the exercise one of characterising such data as well as possible. Clustering and embeddings—such as the one in study III—are two common examples of unsupervised learning. With clustering, we do not have a specific label to assign subjects but rather seek to match subjects in groups with something in common. Figure 6 shows a very simple conceptual example.

Supervised learning

Supervised learning covers classification and regression [74]: we have features and labels, and seek to train a model to make correct predictions for new subjects for whom we only have feature data. More technically, we seek the best possible model *mapping* the features to the outcome [82]. Consider figure 7 illustrating a very

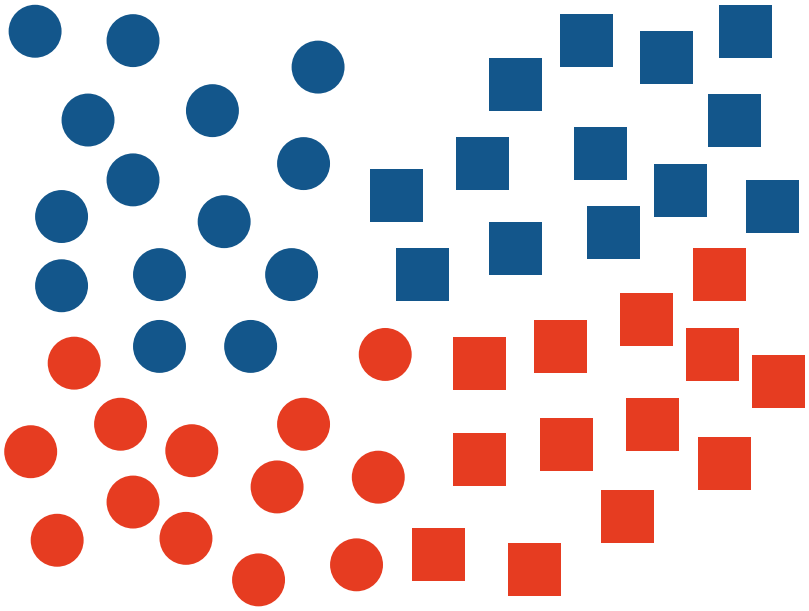


Figure 6: An example of clustering data points. Clustering on colour yields different clusters than by shape. If we consider both shape and colur we get more clusters. With 2 dimensions clustering is not too difficult, but in real-life scenarious with hundreds of dimensions the task becomes much more difficult. Some algorithms find also the best number of clusters, e.g. hierarchical DBSCAN [81].

simplistic classification problem: most would likely agree that the model should predict A = 1, B = 5, C = 8, D = 11.

So, this type of learning is *supervised* because we have a ground truth to learn from, and against which we can compare model performance. More details will follow in Methods.

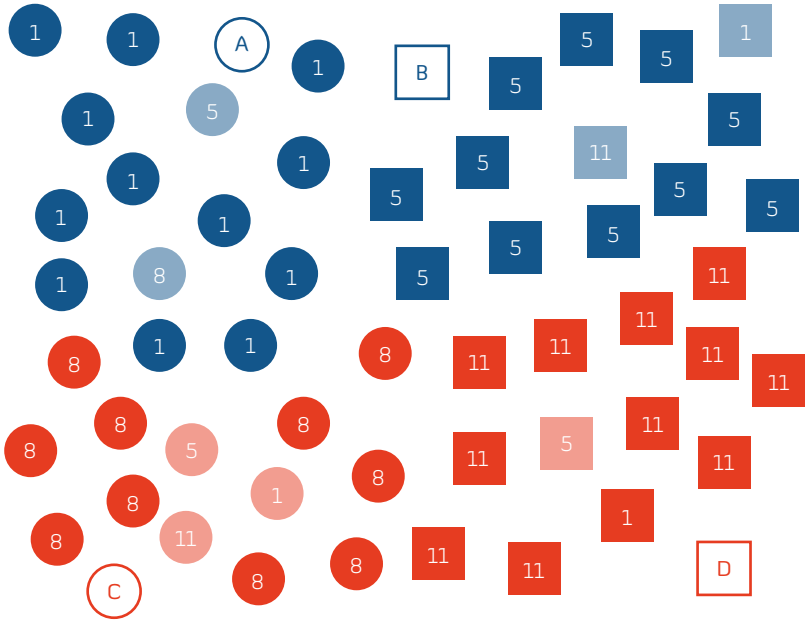


Figure 7: An example of classification of the same data points as in figure 6 but with labels and new observations whose labels we would like to predict.

Data

The Danish registers are well-know internationally for holding rich and linkable data on many axes "from cradle to grave" [83, 84] and have been used by many researchers to answer myriad research questions.

Therefore, in addition to using register data only, the BigTemp-Health research programme collated in-hospital data on all patients at public hospitals in the Capital Region and Region Zealand between 1 January 2006 and 1 July 2016, as detailed below. The two regions comprise approximately 2.6 million citizens, about half the Danish population. These additional data domains include in-hospital medication data, results of biochemical analyses and free text from clinical notes. This heterogeneity of data with long-ranging longitudinal data from national registers and very granular in-hospital data allows for detailed studies of what happens during hospital visits.

This chapter introduces our data sources including their quirks and challenges we faced. The full documentation of the registers is maintained by the Danish Healthcare Data Authority and available (in Danish) online [85].

Chapter contents

The Civil Registration System

Every resident in Denmark has a unique civil registration number (CRN), registered in the National Civil Registration System (CRS). The CRN was put in place in 1968 and now almost all imaginable data are linked to it [83]. A person's CRN consists of their date of birth and 4 arbitrary digits, the last of which indicates their sex. Because the CRNs were scrampled in our data dump to prevent

identification, demographic information—namely sex and date of birth—were extracted from the CRS table with demographic data.

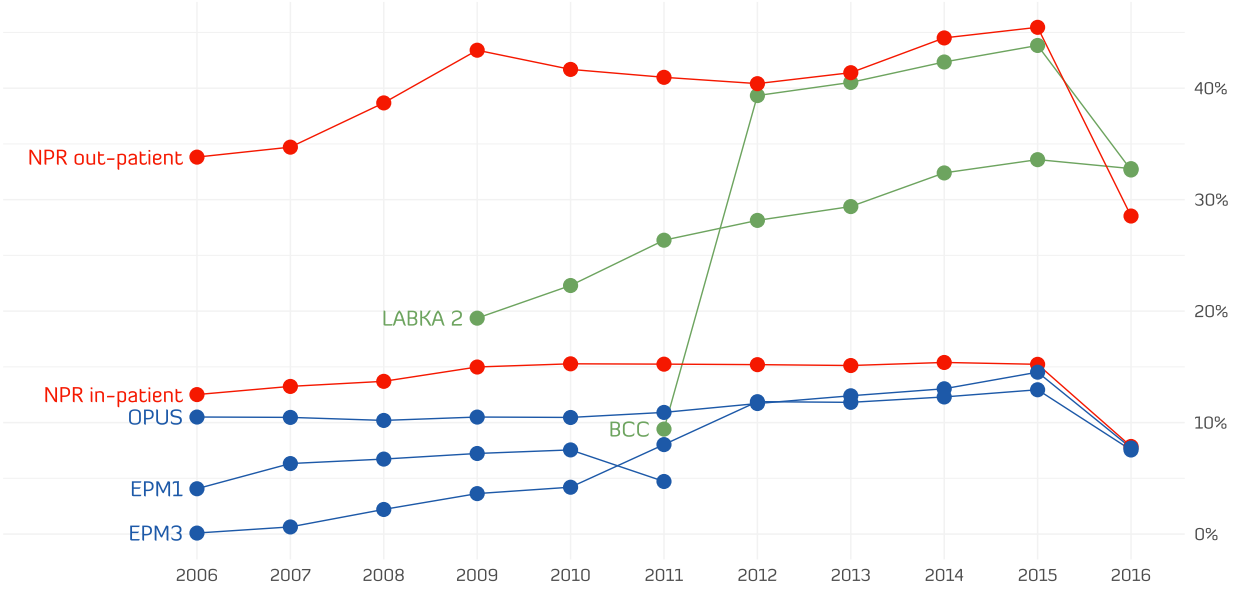


Figure 8: Proportion of full population covered by year and data source. Population sizes computed as the means of quarterly counts [86]. The regions did not exist before 2008, so population counts for 2006 and 2007 set to those of 2008. Red: Diagnoses, procedures, etc. Blue: Medication. Olive: Biochemistry.

The Danish National Patient Register

The Danish National Patient Register (NPR) captures data from a wide range of sources, including clinical quality databases. It is probably the most fundamental register used for observational epidemiological research in Denmark, as it holds various administrative data from all public hospitals,[7] and so makes for an excellent source to characterise patients in terms of e.g. comorbidity and procedures (diagnostic and curative) they have undergone [83]. The register is star-shaped with the main table holding metainformation on each visit at the centre. The temporal resolution differs across subsidiary tables.

Diagnostic data are recorded and timestamped at discharge and so hold no information about at what point during an admission each diagnosis was confirmed. Discharge in the context of the NPR is end-of-stay at the patient's principal department.[8] This entails, for example, that the primary diagnosis is not necessarily the reason for the admission in the first place, and it may not even be reflect the worst condition (from a clinical point of view) for which the patient was treated.

Surgical, therapeutic and diagnostic procedures are timestamped at minute level, on the other hand, and so allow much better for finegrained temporal modelling, both as features and outcomes. Indeed, for example, thanks to the hierarchical structure of the data model (Sygehusvæsenets Klassifikationssystem (SKS), the Classification System of the Healthcare Sector) all procedure codes prefixed by *NABE* pertain to the intensive care unit (ICU) and so

[7] It does capture data from some private clinics and hospitals, but reporting is optional. Due to universal healthcare coverage in Denmark, the private sector is little and the missing data generally of little importance.

[8] *Stamafdeling* in Danish

the presence of any such code is a good proxy for ICU transfer.

A more pharmacovigilance-y application would be using procedure codes to identify what patients need transfusion and when, and use this a proxy for clinical diagnosis of haemorrhage, a known adverse drug reaction (ADR) to certain drugs and drug combinations.

The Causes of Death Register

This register contains information on deaths of Danish residents, for example detailed meta-data on the circumstances, the causal paths leading to death and mode of death [87]. Their validity may be disputed [88], but we only considered all-cause mortality and needed to know the date of death. Date of death, in turn, is likely sufficiently accurate for our needs although some persons are found dead, rendering their dates of death uncertain.

Medication

The medication data arrived in formats clearly designed to underpin production, not research. Collating data from two regions with different information system infrastructures resulted in relatively complex extract-transform-load (ETL) processes to reconcile these data into a sensible, operational format. We only obtained complete medication data for little more than half the BigTempHealth study period i.e. from around 2009 (see figure 8).

The Capital Region data actually came from two sources, Electronic Patient Medication 1 (EPM1) and Electronic Patient Medication 3 (EPM3); their accuracy has been validated previously [89]. EPM1 was rolled out from 2006, explaining[9] the low coverage in the early years. The temporal overlap between the two is an artefact: EPM3 was rolled out starting 2012 (while retiring EPM1), but it seemed that some (historical) EPM1 records made into EPM3, maybe for patients admitted in the early period of the EPM3 rollout.

[9] At least partially: there may have been problems with the actual data dump, but they were beyond repair

For Region Zealand all data were extracted from the OPUS-medicin medication module. An important step in preprocessing the OPUS-medicin data was handling dosing information of individual administrations. The full dose administered was not recorded; instead, we had to compute this (key) data point by combining information on the strength of the formulation given (e.g. tablets with 500 mg. metformin) and the quantity given (e.g. 2 tablets). Thus, we had to ensure meaningful combinations of the units of these values as you cannot administer, for example, 100 ml. of 500 mg. metformin tablets.

In Denmark, medicines are encoded using the Anatomic Therapeutic Chemical classification (ATC), an open-access classification system for drugs using a strictly hierarchical structure for its division of medicines into 5 levels. ATC is maintained by the World

Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology [90] and can be queried online, free of charge.[10]

Due to this tree structure, when used for research one must be careful if exposure to the chemical substance is of interest. If that is the case, we must correctly and mindfully harmonise the exposure to medicines with different ATC codes that are, in fact, the same substance.[11]

One example is ibuprofen. On its own it is classified under 5 ATC codes (C01EB16, G02CC01, M01AE01, M02AA13, R02AX02) and in combinations with codeine (N02AJ08), paracetamol (M01AE51) and oxycodone (N02AJ19). As such, determining exposure to ibuprofen using ATC codes is more involved than it seems at first, especially if you seek to disentangle the isolated effect of ibuprofen. As figure 9 illustrates, this is not an issue for most substances, but two (betamethasone and dexamethasone) have 11 different ATC codes. In addition, diclofenac came out with different results for adverse oucomes in study I (figure 25).

[10] Alternative drug classifications exist, such as WHOdrug (maintained by Uppsala Monitoring Centre (UMC) [91]) and RxNorm [92]. RxNorm is used widely in United States of America (USA), ATC in Europe, and WHOdrug by the industry.

[11] Different strengths alone might be enough for a substance to be classified under different ATC codes [93]

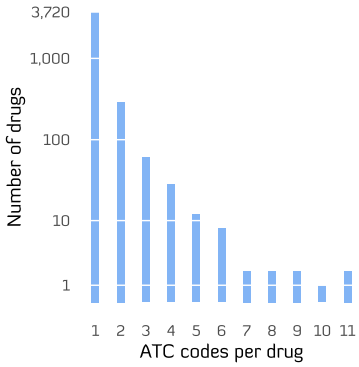


Figure 9: Number of drugs for each ATC code. The bar chart shows that by far most drugs have a single ATC code (3,720 drugs) while few have several ATC codes, and two drugs (betamethasone and dexamethasone) have 11 ATC codes. The vertical scale is pseudo-log-transformed (linear between 0 and 1).

Biochemistry

For Capital Region patients, biochemistry data came from the central database Clinical Laboratory Information System (in Danish: sygehus-Laboratorier, Klinisk Biokemiske Afdelinger) (LABKA), covering the period 2009 through mid-2016, and arrived with a reasonably standardised structure. Our biochemistry data dump from Region Zealand, covering the period 2011 through mid-2016, came from the private vendor of B-Data Clinical Chemistry Laboratory System (BCC) used throughout the region. The raw data contained some 80 tables with sparse or cryptic documentation: after several iterations, the final biochemical results as seen by clinicians were extracted and reconciled with the data from LABKA.

The combined biochemistry data set contained more than 310 million results of samples collected between October 2009 and June 2016, and from 2011 onwards the data set had sufficient coverage in both regions (figure 8). More details on the preprocessing of the biochemistry data will be available in a forthcoming study [94].

In the end, for the studies in this project we used the approximately 13 million estimated glomerular filtration rates (eGFRs) [95, 96].

Clinical notes

Clinical text is a particular kind of natural language with some degree of standardisation but ripe with i.a. ambiguities, domain-specific terminology, non-standard abbreviations, typos, and grammatical inconsistencies. In recent years clinical text is increasingly entered by the healthcare staff themselvs, as opposed to previously when specialised secretaries would transcribe dictated text. The impression is that the heterogeneity has grown as the transcrip-

tion served (also) to somewhat harmonise and proof clinical notes before they were recorded in the system.

Generally, textual data in Danish electronic medical records (EMRs) come in 3 forms. First, *nurses' notes* are usually short (tweet-like in length) with data on the patient's current status, recorded in near-realtime and with relatively high temporal granularity. The frequency of such notes, as well as their content, reflect how intensely the patient is observed. Very ill patients will have frequent notes often with specific observations noted, such as level of consciousness or delirium symptoms.

Second, *written results* of paraclinical examinations such as radiological examinations, histopathological analyses of biopsies and antimicrobial sensitivity evaluations. Such reports usually observe a fairly standardised structure that include relevant paragraphs, sometimes with informative headings.

Third, *clinical notes* fall in two groups: those of inpatients and those of outpatients. Inpatient clinical notes will generally comprise a comprehensive admission note by a junior physician, a thorough review by a more senior physician, several shorter update notes recorded during clinical rounds and a discharge note; there will also be e.g. notes on surgery or invasive diagnostic procedures when relevant. Outpatient notes will be recorded by a mix of physicians, specialised nurses, laboratory technicians, midwives, etc. upon ambulatory visits. These hold focused information on their specific purposes such as post-surgery check-ups, controlling pacemaker status, and regular pregnancy visits.

We had access to approximately 75 million notes of the last type. Some were clearly form-like with large special-character/any-character ratios. Automatically extracting information from such notes is usually impractical because the information is ambiguous. For example, it is not uncommon to see formats akin to these:

```
GA [34+2] Headache No [ ] Yes [x] Previous c-section
No [x] Yes [ ]

Gestational age: 34+2. Headache __ yes. Previous
c-section no ___.
```

It requires little medical knowledge (and exposure to obstetrical patients) to accurately extract the information held by these two examples; recognising also that they hold the same data is a easy, despite GA being a very common abbreviation for *general anaesthesia*.

Enabling a computer to do this automatically, in a way that scales and correctly handles myriad non-standard encodings, is much more difficult. Even manually annotating a large corpus of form-like notes does not guarantee satisfactory results.

As an aside, from a natural language processing (NLP) viewpoint the latter of the two examples is easier, because it actually holds only the information of interest whereas the first example

has both options from which the healtchare professional could (and did) choose.

Methods

This chapter details the main methods of our studies, following the logic illustrated in figure 5; word embeddings are presented first (although used only in study III) as they serve to operationalise the textual data for analysis.

Chapter contents

Word embeddings

Clinical notes usually arise from very structured data collection, but this structuredness is lost when saved as free text without formatting or terminology conventions. Free text is rich because different persons, in their own words, can describe and record their observations, interpretations, and conclusions. This strength, however, also complicates matters as such data are unfit for quantitative analyses; natural language processing (NLP) tools enable us to transform textual data to recoup some of the structure and make them compatible with i.a. statistical models [69].

The exact nature of such data transformation depends on the purpose of the analysis and can take many forms: from very sophisticated semantic analyses yielding relationships between words, over more brute force approaches assigning codes to them based on (fuzzy) matching against ontologies[12], to converting words (or sentences or entire documents) into numeric vectors.

The vector approach typically comes in two variants: one-hot-encoding and word embeddings. Traditional one-hot-encoding represents each word by an N-dimensional vector where N is the number distinct words in the corpus.[13] This creates huge vectors with little information due their extreme sparsity: a single cell will hold the value 1, the rest will be 0.

[12] Such as the 10th revision of the International Classification of Disease (ICD10) or the Medical Dictionary for Regulatory Activities (MedDRA)

[13] The corpus is the collection of e.g. documents or clinical notes from which we extract information [97, ch. 1]

This has a major downside, other than the computational inconvenience: even words with similar meanings or that contain little typos are considered as different as two completely unrelated words. This happens because the N one-hot vectors form the standard basis for $\mathbb{R}^N$ (the vector space into which we have mapped the words) and are, thus, by definition linearly independent [98, ch. 4].

Word embeddings overcome these disadvantages by packing text into much, much fewer dimensions, usually a few hundred. Embedding models are often trained in an unsupervised manner although they can be incorporated as submodels in larger prediction models [70].

In essence, learning the embedding constitutes dispersing the words of the corpus in the embedding space so that words that represent similar notions are closer to each other than to other words. To concretise this, consider the 11 words[14] below with their one-hot vectors:

[14] COPD: chronic obstructive pulmonary disease. NIV: non-invasive ventilation.

$$
\begin{aligned}
\text{tonsilitis} &= \begin{pmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{neuralgia} &= \begin{pmatrix}0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{headche} &= \begin{pmatrix}0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{sore throat} &= \begin{pmatrix}0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{impotence} &= \begin{pmatrix}0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{pneumonia} &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{NIV} &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\end{pmatrix}\\
\text{man} &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0\end{pmatrix}\\
\text{COPD} &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\end{pmatrix}\\
\text{dialysis} &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0\end{pmatrix}\\
\text{kidney failure} &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1\end{pmatrix}
\end{aligned}
$$

Instead of representing each word with an 11-dimensional one-hot vector (as above), the embedding in figure 10 represents each word by a 2-dimensional dense vector. Using a 2-dimensional embedding helps us build intuition that holds even when the embedding space has (many) more dimensions and, thus, cannot easily be drawn on a piece of paper.

The black arrows show the embedding vectors starting in $(0, 0)$: words that are closer in the embedding space have similar vectors. Fascinatingly, word embeddings enable language algebra[15] with a conceptual example being:

[15] A classic example is that *england* + *london* − *france* = *paris* [69]

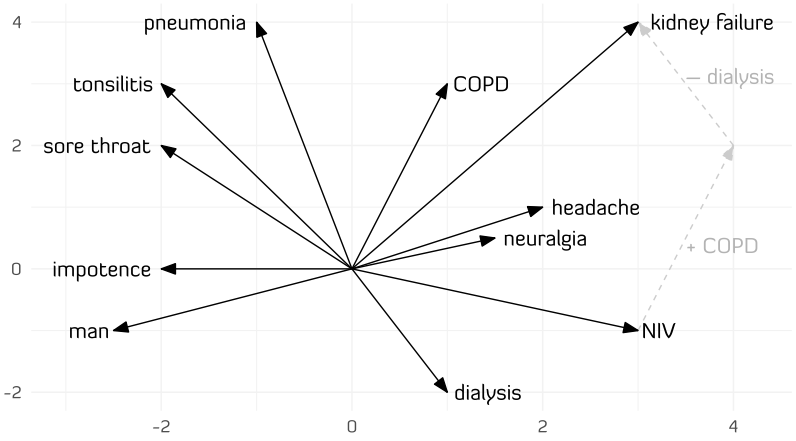


Figure 10: Simple example of embedding 11 words into a 2-dimensional embedding space. For example, the embedding vector of tonsilitis (-2, 3) and for sore throat (-2, 2) are very different from those of dialysis (1, -2) and headache (2, 1). The grey +COPD vector is identical to the black COPD one but moved to the end of the NIV vector; the grey –dialysis vector is a reversed version of the black dialysis vector and moved to the end of the grey +COPD vector to arrive at kidney failure.

$$\begin{aligned}\text{NIV} + \text{COPD} - \text{dialysis} &= \begin{pmatrix}3\\-1\end{pmatrix} + \begin{pmatrix}1\\3\end{pmatrix} - \begin{pmatrix}1\\-2\end{pmatrix} \\ &= \begin{pmatrix}3\\4\end{pmatrix} = \text{kidney failure}\end{aligned}$$

While several ways exist to quantify similarity between vectors, the most common is arguably cosine similarity i.e. the angle between them. Figure 11 shows all pairwise cosine similarities in the corpus using embedding and one-hot vectors, respectively. Two things are evident: when using word embeddings, "clusters" of clinically related words light up (blue) and one-hot vectors are completely dissimilar.

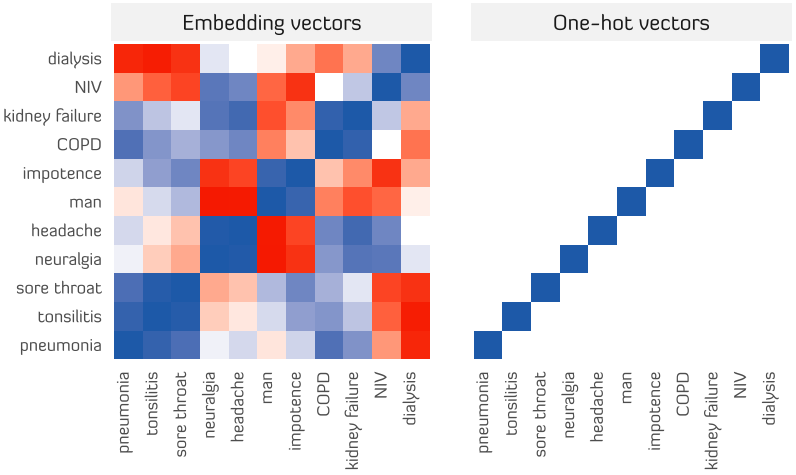


Figure 11: Cosine similarities between the words in figure 10, using embedding and one-hot vectors. The similarity scale goes from -1 (red, completely opposite meanings) through 1 (blue, exact same meaning).

The property that words with similar meanings, regardless of spelling, yield similar embedding vectors makes word embeddings a powerful tool for operationalising free-text data for machine learning models and have been put to use in a variety of ways. So-

called entity embedding builds on the same notion but also embeds non-textual data [99].

Process modelling

Once data are made suitable for statistical analysis, we use them to model the data-generating process: the hidden process in the world from which the data arose in the first place.

Generalised linear models

Generalised linear models (GLMs) have long been the mainstay for effect-size estimation with observational data or even in sub-group analyses of randomised controlled trials (RCTs) because their coefficients have convenient interpretations. We can model biological interactions or extend the models to include non-linear effects with e.g. splines, at which point the models are called generalised additive models (GAMs) because they no longer assume and model only linear effects.

Devising sound GLMs requires domain knowledge and evidence from previous research or justifiable assumptions about causal assocations (or lack thereof) between included variables in these models. This work is tedious, and much effort has been put into yielding causal estimates while bypassing this step [100–103].

GLMs are actually the extension of linear models to handle other outcomes than unbound, continuous outcomes. This is achived with a so-called *link* function g (more on this in Outcome modelling):

$$g(\mathrm{E}[\mathbf{Y}|\mathbf{X}]) = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p \quad (1)$$

$$= \mathbf{X}\boldsymbol{\beta} \quad (2)$$

$$\mathrm{E}[\mathbf{Y}|\mathbf{X}] = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \quad (3)$$

GLMs are sometimes called *shallow learners* (as opposed to *deep learners*) because they connect each feature (and interaction term when present) directly with the output. Thus, a basic vanilla GLM model assumes that each feature has a linear effect on the transformed outcome and that this effect be independent of the effects of the other features. This yields directly interpretable parameter estimates (e.g. odds ratios for logistic regression models) and, as a consequence, the possibility to estimate average marginal or conditional effects.

This assumption of independent effects is often violated or at least biologically unlikely. Consider, for example, predicting the risk of 30-day mortality as a function of age, sex and the need for intensive care. The impact of needing intensive care is probably greater in older patients than in younger: age and intensive care need probably interact. Unless explicated in the GLM, such relationships will evade the model's learning. Failing to capture such a

crucial characteristic of the data-generating process will likely result in biased effect-size estimates (aetiologic enquiries) or suboptimal model performance (predictive enquiries).

These assumptions are especially important when the interest is in the coefficients, because they estimate the relative effect of unit-changes in each of the features on the transformed outcome. Assessing these assumptions becomes difficult if interaction terms or even splines are included.

Penalised regression models

GLMs can work well when the number of observations N is greater than the number of parameters k in the model. As k approaches (or supersedes) N, GLMs become increasingly problematic due to over-parameterisation. This enables the model to fit the developmet set very well (perhaps even identify single observations) but impedes generalisability [104].

One solution is to add shrinkage to the coefficients by penalising model complexity, forcing the model pick as much information as possible from as few features as possible. Thus, by adding so-called regularisation terms to equation (2) we obtain:

$$g(\mathrm{E}[\mathbf{Y}|\mathbf{X}]) = \mathbf{X}\boldsymbol{\beta} + \lambda|\boldsymbol{\beta}| + \phi\|\boldsymbol{\beta}\|^2, \tag{4}$$

where $|\boldsymbol{\beta}| = \sum_{i=1}^{k} |\beta_i|$ is the ℓ_1 norm and $\|\boldsymbol{\beta}\|^2 = \sum_{i=1}^{k} \beta_i^2$ is the ℓ_2 norm. Equation (4) offers three types of penalised regression models: lasso regression when $\phi = 0$, ridge regression when $\lambda = 0$, and elastic-net regression when $\lambda, \phi > 0$ [104]. λ and ϕ are so-called hyperparameters that can either be pre-specified or learnt with the coefficients (see Hyperparameters and how to tune them).

Lasso regression pushes as many coefficients as possible all the way to zero resulting in feature selection. Ridge regression, on the other hand, will generally only push coefficients *toward* zero enabling i.a. better handling of collinearity because it need not pick which to keep. A mix of the lasso and the ridge, the elastic-net regression will do some feature selection but do so more leniently.

We again consider a 2-dimensional example to build intuition (figure 12). The lasso estimate must lie within the square, and the ridge estimate within the circle.[16] For both, the estimate will be combination of coefficients that yield the highest likelihood within these constraints, represented by the oval lines.

Smaller values of λ and ϕ will yield larger squares and circles: as λ and ϕ tend to 0, equation (4) tends to equation (2). That is, larger values of λ and ϕ will result in more shrinkage and stronger feature selection (for the lasso).

[16] This follows from the squared-error loss function (not shown); because they are simpler, the example in figure 12 is actually based on a linear model

Multilayer perceptrons

One way to overcome manually specifying complex models while learning ditto relationships in data is using multilayer perceptrons

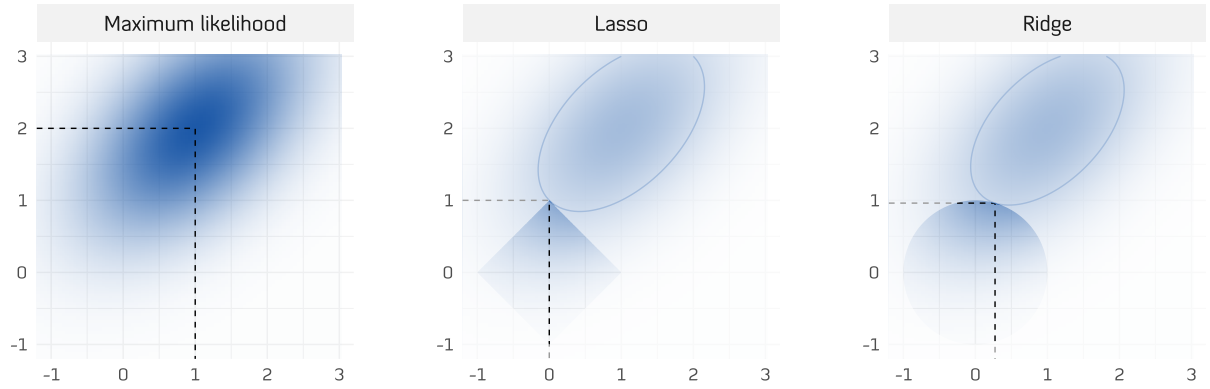


Figure 12: The contour of the 2-coefficient likelihood function with the constraints imposed by the lasso and the ridge. The dashed lines indicate the maximum likelihood (left, $(1, 2)$), lasso (mid, $(0, 1)$) and ridge (right, $(0.27, 0.96)$) estimates. The oval cirles in the lasso and ridge sub-plots show the contour line of the estimates. The lasso shrunk one parameter all the way to 0 while the ridge parameter estimate for the same parameter is 0.27. Adapted from figure 16.4 in Efron and Hastie [104].

(MLPs), a basic type of artificial neural networks. Although the brain was the inspiration of the name and architecture of such models, they do not truly mimic the brain and probably provide much less computing power than neurons [105]. MLPs have one or more hidden layers each with a number of nodes. MLPs with several hidden layers, as well as more complex neural network models with specialised architectures, are often collectively called *deep learning* models, but since we only used MLPs, discussing other types of neural network models is beyond the scope of this section.

MLPs allow each feature to play a signifant role in many directions of the prediction because each hidden layer contains several nodes, and each feature can affect each hidden node independently of its influence on the other nodes. Thus, even a single hidden layer adds much flexibility compared with the GLM in which each feature influences the prediction through a single coefficient.[17] As we add more hidden layers, this complexity increases somewhat exponentially and the MLP, consequently, can capture highly non-linear effects of single features and complex interactions between them.

[17] Indeed, an MLP with a single hidden layer can approximate any function arbitrarily well if that layer has enough nodes [75, sec. 6.4]

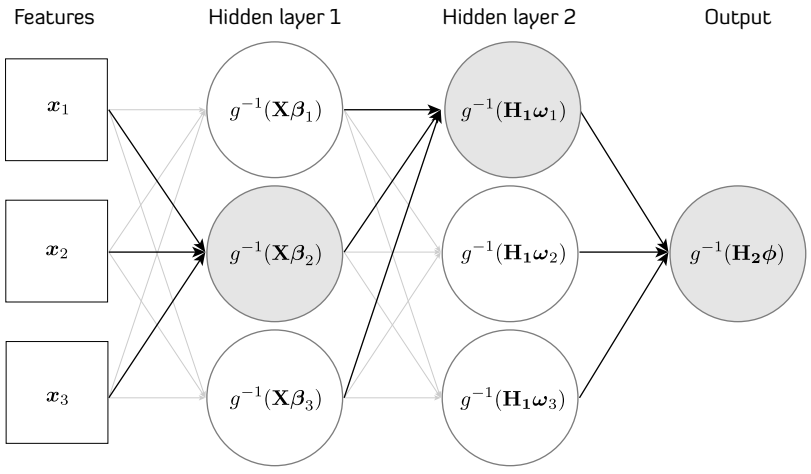


Figure 13: Multi-layer perceptrons are multiple stacked generalised linear models. One node per layer highlighted solely to illustrate this notion. Adapted from i.a. Thorsen-Meyer [106]

It is important to remember, however, that MLPs come about from a series of *linear* transformations and so, these models can only capture complex relationships insofar as they be amenable to such serial linear transformations [70, 107].

Thus, MLPs models learn ("automagically") the structure needed to predict well and seem to alleviate the model-design burden as we make no assumptions about the data-generating process. There is, of course, no free lunch and the challenges regarding model design and interpretation are discussed later.

Similar to the link function in GLMs, each node in the hidden layers of a MLP applies a so-called activation function that serve to map the outputs of all nodes in the same hidden layer into the same range of values to prevent a single node from dominating that layer's output.

The logistic function is amongst the most common activation functions, and an MLP using logistic activation functions essentially becomes a compound model of (potentially many) stacked logistic regressions. Indeed, the notation in figure 13 is chosen to emphasise this key insight: the equations inside the nodes in the hidden layers and the output node are equivalent to that of the GLM in equation (2). In GLMs the learnt parameters are usually called *coefficients* whereas they are called *weights* in neural-network models.

Outcome modelling

Apt operationalisation of the outcome variable is crucial no matter whether the model serves a predictive or aetiological purpose. Outcome data types can, generally, be divided into continuous and categorical. In the absense of comparable observation time, a temporal aspect of the outcome can be incorporated as in time-to-event or per-unit outcomes. Many subtypes exist, but in this thesis I have used one from each: binary (i.e. the occurrence or not of an event if interest) and count (length-of-stay, in study I).

Multivariate outcomes are probably more common in machine learning than epidemiology at large, perhaps because the models are already so high-dimensional and interpretability of the individual variables less important, so adding more complexity usually comes with only slight marginal costs.

Logistic regression

Clinical decisions are often binary: you operate or you do not, you admit the patient or you do not, you give the patient the drug or you do not. Although ultimately binary, these decisions result from intricate decision-making processes, in which predictions (conscious or not) of the patient's risk of harm in either case plays a crucial role.

We can directly observe and model continuous outcomes such as

In machine learning, binary-outcome models are usually called *classification models* because their application tends to focus on assigning labels to persons, Twitter posts, images, etc. In epidemiology this term is somewhat misguided: these models need not report a label assignment but mostly yield the probability of belonging to each of the possible outcome groups. Thus, we only arrive at a binary decision once we set a threshold for the probability of event. Thus, I instead prefer and use the term *binary regression* [108].

body weight, length-of-stay, and number of admissions. In contrast, with a binary outcome (e.g. survival or not) we are interested in the probability of experiencing the event, but we cannot observe this probability directly. Instead, in a given cohort after a certain amount of follow-up time has passed, we want to estimate the proportion of subjects with a certain combination of features who survived (in this example).

That is, we seek to estimate the expected value of a Bernoulli process (i.e. the probability of observing the event $\mathbf{Y}_i$) for a given combination of feature values $\mathbf{X}_i$ [74]:

$$y_i \sim \text{Bernoulli}(p_i) \tag{5}$$
$$p_i = E[\mathbf{Y}_i \mid \mathbf{X}_i] \tag{6}$$

With few predictors we can do this with stratification, but even a modest number of features renders stratification impractical and it will not provide estimates for feature-value combinations not seen in the data set. Logistic regression remedies these shortcomings.

Because probabilities can be neither less than 0% nor greater than 100%, a convenivent transformation is to model the logarithm of the odds of observing the event. This is called the logit of the probability[18] and is convenient for at least two reasons: the unbounded codomain (see figure 14) ensures that the transformed outcome is appropriate for any linear combination of features, and there is no lower or upper limit to the relative change in odds.[19]

[18] Log-binomial regression directly models the risk but comes with its own set of complications [64, 109, 110]

[19] A probability can not necessarily be doubled; for example, the maximum relative change of a 80% mortality is 25%

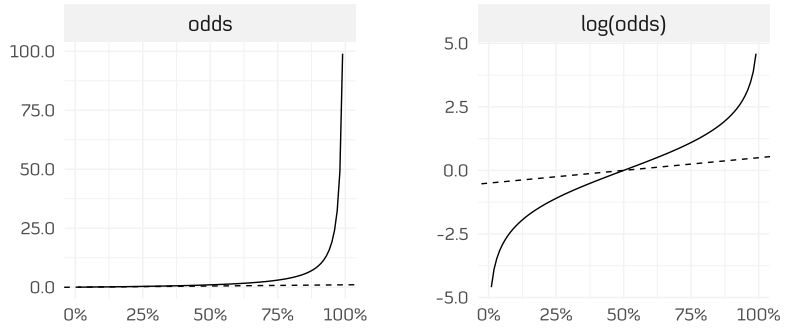


Figure 14: The logit is defined as $\text{logit}(p) = \log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$ and maps the probability p from $[0, 1]$ to $(-\infty, \infty)$. The dashed lines show the 1:1 relationships for comparison.

The logistic regression model is a particularly popular GLM for modelling binary outcomes and the baseline approach for binary classification in machine learning. The name of logistic regression arises from the fact that the *logistic* function is the inverse of the logit, so that our model actually is on the form,

$$p_i = \text{logit}^{-1}(f(\mathbf{X}_i)) = \frac{1}{1 + e^{-f(\mathbf{X}_i)}}, \tag{7}$$

where f is the function, be it linear or not. As such, in a logistic regression with a GLM setup, the coefficients are the odds ratios

for a unit change in the feature value, a decent approximation to the risk ratio as long as the event probability be low. In machine learning setups, we usually employ slight variations of the classic logistic regression to model the probability directly.

Standardised difference in proportions

When no stratification or other adjustment for confounding is needed, the standardised difference in proportions (SPD) is a good metric for comparing binary univariate outcomes in two populations. SPD is defined as

$$d = \frac{P_A - P_B}{\sqrt{\frac{1}{2}P_A(1-P_A) + \frac{1}{2}P_B(1-P_B)}} \cdot 100\%, \tag{8}$$

where P_A and P_B are the observed proportions of events in the two groups. $d > 0$ if the event proportion is greater in group A than that in group B and vice versa.

SPD is an extension of the standardised difference in means (which is valid only for continuous variates) and is convenient because its value carries information about both the strength of association and its significance: a SPD of 10% corresponds roughly to a p-value of 0.05 [111]. SPD only yields a point estimate, but confidence intervals can be computed with e.g. bootstrapping [104, 112].

Poisson regression

A regular linear regression may be inappropriate for counted outcomes because it allows negative and decimal outcomes. In counted outcomes such as length-of-stay, negative outcomes are impossible and often we count full days or choose another scale with appropriate, integer values (e.g. length-of-stay in hours instead of full days). Poisson regression is one appropriate approach for such data.

A basic Poisson regression is relatively simplistic, assuming that the variance equal the mean of the Poisson distribution; there are ways to handle situations when this assumption is violated [109] such as introducing a parameter to capture the dispersion. Further, Poisson regressions can be offset to account for, for example, varying observation or exposure periods to effectively model the number of events per some index unit (e.g. per day, per patient or per department) [109].

Poisson regression models are also called log-linear models because the log-transformed outcome variable is modelled via a standard linear model [109]. More generally, we can plug in any model, be it linear or not, to model the λ parameter of the Poisson distribution,

$$y_i \sim \text{Poisson}(\lambda_i) \tag{9}$$

$$\log(\lambda_i) = f(\mathbf{X}_i) \tag{10}$$

Even though the mean of a Poisson distribution must be greater than zero, zero-count observations are possible (especially for small λ), see figure 15. When the data contain excessive zeros, rather than being over- or underdispersed, one can invoke i.a. hurdle or zero-inflated models [113, 114].

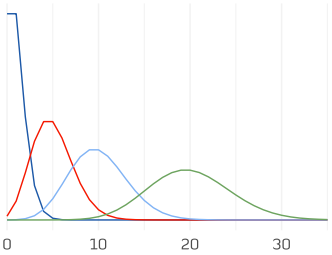


Figure 15: Poisson probability density functions for each of 4 different values (1, 5, 10, 20) of λ.

Cox regression

The outcomes described above are univariate, but often in realistic settings univariate outcomes do not suffice. When the outcome of interest is binary but follow-up time is not (approximately) the same in all patients or not all patients can be expected to experience the event before end of follow-up, we combine the outcome indicator with a variable that specifies the time-to-event (or censoring). Figure 16 illustrates a simple example of time-to-event data: binary outcome with some censoring, that is, patients in whom the outcome is not observed.

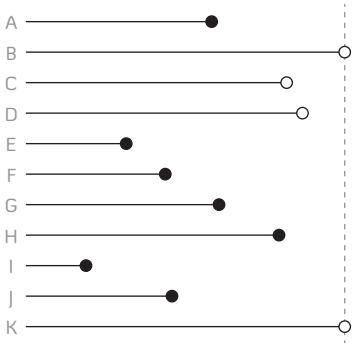


Figure 16: Time-to-event observations in 10 fictive patients. Filled circles represent the 7 patients experiencing the event in the follow-up period, hollow circles those who do not. Patient B is censored at end of follow-up; patients C and D are lost to follow-up. Inspired by i.a. Altman [115, fig. 13.1].

The censoring of patients who did not experience the event before end of follow-up is often assumed non-informative. Censoring during the follow-up period, on the other hand, can be problematic with the notable exception of patients included, say, 3 months before the end of a study with a 6-month follow-up period.

Cox regressions, or proportional hazards models, seek to estimate the hazard at a certain time t, making no assumptions about the distribution of the baseline hazard h_0 [115, 116],

$$h(t_i|\mathbf{X}_i) = h_0(t)\exp(\mathbf{X}_i\boldsymbol{\beta}) \tag{11}$$

This setup is convenient for effect-size estimation because the baseline hazard h_0 does not matter.

Training multi-layer perceptrons

Design decisions abound when crafting MLPs. At the most general level, training the model means finding the parameters and hyperparameters (see below) that yield the best fit, as with any machine learning model type. Parameters are learnt minimising the loss (corresponding to maximising the "goodness-of-fit") through so-called gradient descent [70, 104]. We do not (generally) know the shape of the loss function, so we must chart it through an iterative process to hopefully find the global minimum where we obtain the best possible fit.

Every step in this iterative process is called an epoch, and we illustrate the training process with so-called learning curves. Crucially, training could go on forever and the model might continue

to reduce the loss, although increasingly slow until some plateau is reached (figure 17). This "saturation" often causes over-fitting: the model will learn characeristics specific to the training data and will not transport well to new data, see Avoiding over-fitting.

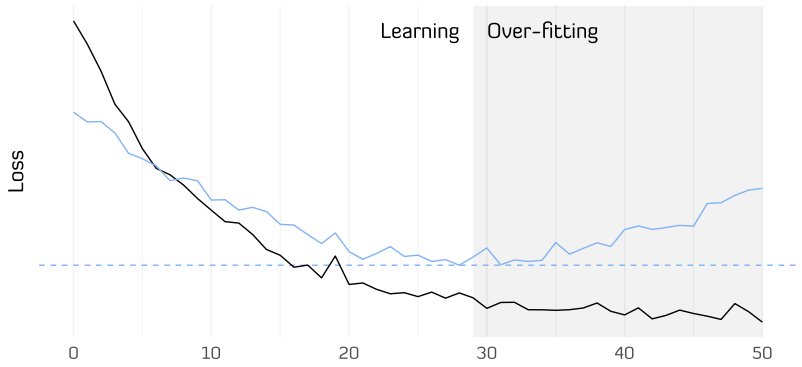


Figure 17: Examplary learning curve for training (black) and validation (blue) over 50 epochs (x axis; zeroth epoch = the random initial parameter values). The loss (arbitrary scale, so hidden) shown on the vertical axis. The blue dashes line shows the best epoch in the validation set.

In a machine learning context, we operate with several names for data sets depending on what purpose they serve. We generally distinguish the development set from the test set; the test set is, somewhat confusingly, used for validation and can be internal or external [117]. When training the prediction model, the development set is usually split into a training set and a validation set; when employing e.g. 5-fold cross-validation the development set is actually split into 5 training-validation set pairs [118, ch. 17].

Avoiding over-fitting

It is not too difficult to devise an MLP that fits the development set excellently. The tricky part is to stop training the model when it has learnt as much as it can from the development data while still transporting well to other data, increasing its utility in any target population. That is, to avoid over-fitting. This tradeoff is illustrated in figure 18.

In the ideal scenario we have low bias and low variance, so that we are quite convinced about our result, and that result is quite accurate. When we continue training the model in the training set, the variance will go down but its bias (in the validation set) will go up, at which point we begin overfitting.

Thus, regularisation seeks to prevent over-fitting by striking the best possible bias-variance balance [119] and comes in a variety of approaches, e.g. using parameter norm penalties (just as in Penalised regression models) and early stopping [75, ch. 7]. Early stopping has the obvious advantage that we explicitly stop training the model when performance in the validation set deteriorates, but early stopping hinges on this very validation set, and so the model needs to be trained on the full development set afterwards in a way that utilises the information gained from the early stopping [75, sec. 7.8].

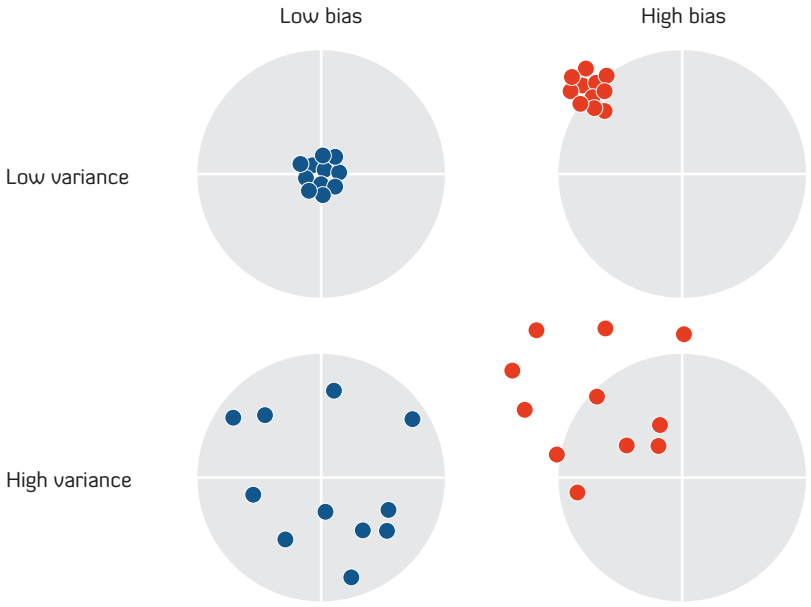


Figure 18: Classic illustration of the bias-variance tradeoff. The ideal in the upper-left corner represents high certainty about the estimate that is close to the truth; the worst situation is the bottom-right corner with much uncertainty about the estimate that is far from the truth.

Hyperparameters and how to tune them

In essence, a statistical model has two types of parameters: learnt parameters and hyperparameters, and we seek to find the best-performing combination of these that still transports well. Learnt parameters are updated during training (it is the whole point of training the model) to make the model fit the data increasingly well: coefficents are learnt parameters in GLMs, so are weights in MLPs.

In contrast, any given statistical model rests on a number of design decisions about its architecture, reflected in the model's hyperparameters that set the stage for and remain fixed throughout training. These can be crucial to the model's ability to learn pertinent relationships from the data. A key strength (and selling point) of machine learning models is that you can take the human almost entirely out of the training process and automate many decisions, not least with respect to hyperparameters.

A logistic regression model, for example, has ony few hyperparameters such as the link function. In lasso and ridge regression models, the penalties are hyperparameters. MLPs have many more hyperparameters, including the number of hidden layers, the number of nodes in each layer, and activation function(s). While there are ways to make meaningful, conscious decisions about the link function in a binary-outcome GLM, it is difficult to know a priori what penalty is the best in a ridge regression or how many layers yield the best performance in a deep learning model.

One remedy is automatic hyperparameter optimisation whose simplest approach is an exhaustive grid search over all combinations of the hyperparameter values. Although conceptually very simple, it comes with two serious disadvantages. First, combinatorial explosion might render this approach impractical: 4 hyperpa-

rameters with each 5 values yields 1,024 different combinations so if training a single model takes 15 minutes, finding the best hyperparameter combination would take 21 days. Second, continuous hyperparametres (such as norm penalty) must be binned to play well with grid searches.

Instead of exhaustive searches, one can take a number of random samples from the hyperparameter space, train the corresponding models, and use the one that comes out best. A conceptually and empirically more appealing way is to build some form of systematism into the sampling to optimise the hyperparameters instead of blindly sampling random values. This is what we did in study II (see figures S3–S12), using Optuna's multivariate *TPE sampler* [120] based on Bayesian optimisation [121].

Evaluating prediction models

The evaluation, importance and interpretation of model performance depends heavily on the purpose of the model. Our machine learning models were all predictive and so here I outline two crucial aspects of model performance evaluation (discriminatory ability and calibration) and a relatively new approach to evaluating the potential clinical utility (decision-curve analysis). Performance should be assessed in test sets, be they internal og external.

Validation

At its core, validation is about quantifying to what extent the model makes good predictions. Keeping the target population in mind is important: the evaluation should gauge how well we expect our results to transport to the target population and not just any random population from anywhere in the world[20] [106]. The decision-making processes we model greatly depend on cultural and structural conditions far beyond what is captured in our (often, clinical and demographic) data [80]; these structural conditions are likely so pervasive that even if model performance in an external data set from another country is moderate, it will not necessarily give any useful insights into whether the model would benefit our future patients in that setting [122].

[20] Unlike causal relationships we would expect to apply (more) universally

There are several ways to split the original data set into development, validation and test sets. Here, I briefly describe the types and variations we employed in studies II and III. First, split-sample validation is probably the most common scheme in machine learning applications and very simple: the dataset is split randomly into two disjoint subsets (usually with a 4-to-1 ratio = 80%-20% split-sampling). This scheme makes a quite strong assumption of a stationary data generating process, so that any patterns learnt from past data will apply equally in future data. This is unlikely to hold and split-sampling validation is by some considered to "only work when not needed" [117, p. 245]. For cross-validation as part of early

stopping, we found this a viable approach despite its limitations.

Second, temporal validation better resembles the overall purpose of building prediction models: we seek to make predictions about the future using data from the past. Even if the model learns to predict outcomes very well in the development set (yielding *apparent performance*) once we make predictions about the future, these will likely be off-base to some extent (*real performance*) due to drift in the underlying data-generating process. Even when conducting external validation, if the validation and development data were collected concurrently, the temporal drift (e.g. changes in clinical guidelines or novel medical technologies) will likely escape the validation scheme.

Third, K-fold cross validation in its most basic version randomly divides the development set into K subsets (called *folds*) of equal size. It then uses $K-1$ folds for training and the last fold for validation and does this K times to obtain K estimates of the prediction error [104, ch. 12].

Fourth, group K-fold cross-validation is appropriate when the units of analysis are not independent. This happens, for example, if we include all admissions and any patient can (but need not) contribute several admissions to the dataset. If this is not accounted for, a patient may appear in both the training and validation set, potentially leaking information between them and allowing the use of validation data during training [70].

Finally, stratified K-fold cross-validation is used for prediction models with categorical outcomes and ensures that the distribution of the outcomes in the training and validation sets be approximately the same as in the full development set.

Discrimination

A model's ability to distinguish patients with the event from those without reflects its discriminatory power. Models with better discrimination have greater variance in the predictions [123, 124], allowing for better separation of event and no-event individuals. Discrimination in binary regression models is usually visualised with the receiver operating characteristic (ROC) curve and quantified by area under the ROC curve (AUROC) [118]. AUROC has the desirable property that it accounts for the trade-off between specificity and sensitivity across cutoff values; other metrics, such as Matthew's correlation coefficient (MCC) that is more robust to (somewhat extreme) imbalances in the outcome variable [125], are evaluated at a specific cutoff as it relies on a single instantiation of the confusion matrix derived from the predictions.

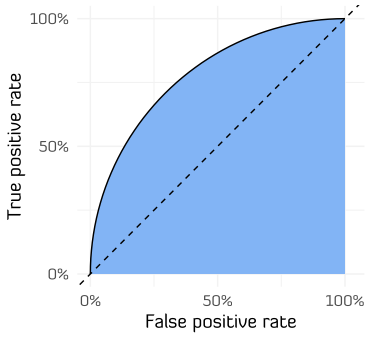


Figure 19: Receiver operating characteristic curve.

Calibration

Calibration gauges to what extent the actual predictions be accurate. A poorly-calibrated model yields incorrect predicted risks which can hamper its utility: guiding patients' (and clinicians') decision-

making on whether to undergo (perform) a procedure crucially depends on accurate risk predictions; a believed 5% mortality risk could sway towards the procedure whereas knowing the real risk to be 25% might cause the patient to forego it. Calibration alone is insufficient as a performance metric because simply predicting the incidence for all individuals will have perfect calibration [123] but will be of zero clinical use.

We used calibration-in-the-small throughout, plotting decile-binned predictions on the x axis and bin-wise event proportions on the y axis. Decile-binning is usually used; in a recent study we found no qualitative difference of using 15 og 20 bins [126]. The final step is to fit a linear regression to these coordinates. A fitted linear line that follows the diagonal represents perfect calibration.

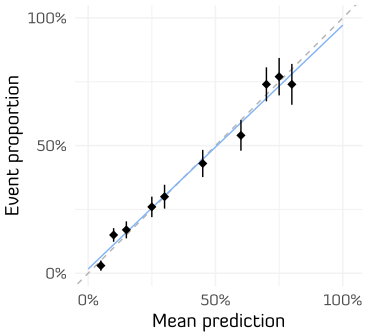


Figure 20: Calibration curve with fitted linear regression.

Decision-curve analysis

Discrimation and calibration together thoroughly gauge performance from a purely technical perspective, but they fail to evaluate the model with respect to clinical utility: will we make better decisions with the model than without it [127, 128]. To be sure, assessing true clinical utility is different and involves many aspects, not least financial and organisational, but decision-curve analysis provides a sound framework for obtaining some insights into the potential clinical utility of a model, with certain assumptions [127].

Decision curve analysis is visual using overlain curves representing different actions. The most basic plot has three such curves: an *intervene-in-all* curve, an *intervene-in-none* curve, and an *intervene-in-flagged* curve where flagged patients are those predicted to be at high risk by the model. The x axis is the prevalence of the event in the target population, the y axis is an arbitrary net-benefit scale with no immediate interpretation, but one can compare directly the net-benefit (which can be positive or negative) of the three actions.

Crucially, the plot cannot be used to pick a threshold: the x axis reflects the prevalence one expects in the target population, and so one should find the corresponding value and identify the decision curve with the more favourable net-benefit. If this is the *intervene-in-flagged* curve, the model can be considered to hold potential clinical utility, subject to verification in a prospective study.

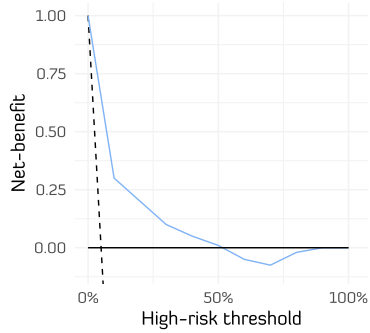


Figure 21: Decision curve. The intervene-in-flagged curve (blue) suggest that this fictive prediction model be of clinical utility.

Explaining predictions

Put simply, in aetiological epidemiology a statistical model should perform well because this adds confidence in its effect-size estimates, in turn derived from the learnt parameters; in contrast, the merits of a prediction model hinges on its ability to make accurate predictions, and the learnt parameter estimates are simply a means to this end.

Consequently, machine learning models are built to learn complex relationships somewhat autonomously without explicit pro-

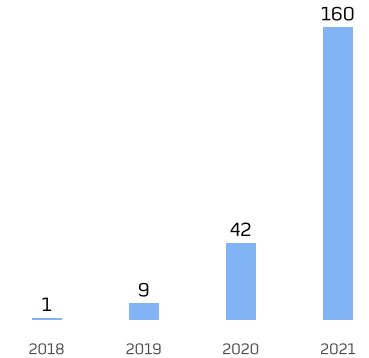


Figure 22: Number of hits on PubMed (y axis) by year (x axis). Query on 18 October 2021: `"shap value*" OR (shap AND (explain* OR explanation*)) OR "shapley additive explanation*"`.

gramming, with the underlying assumption is that there is no assumption: that we "learn from the data" and that latent (= hidden) relationships can be brought fourth "automagically" insofar as we have enough data. Learning extremely complex relationships precisely underpins the goal (accurate predictions) and we do not care for their interpretations from an effect-size viewpoint. Consequently, MLP models do not innately lend themselves well to scrutiny, and so we need other techniques to gauge the plausibility of the relationships they have learnt.

The growing recognition that entirely opague predictions in the medical field are of little use and may hamper uptake [129] has given rise to the discipline of explainable artificial intelligence (xAI) that seeks to alleviate this important shortcoming of complex prediction models [73, 102, 130–132].

We used one of the available methods for explaining predictions, the SHapley Additive exPlanation (SHAP) framework [73], gaining traction in the medical field (figure 22). The method yields one so-called SHAP value per feature per unit of analysis. For binary regression, the SHAP value is the absolute change in risk of a given unit's value for each feature. Put simply, if you take the mean risk across all units[21] in the cohort and add the sum of one unit's SHAP values, you arrive at the predicted risk of that unit, as illustrated in figure 23.

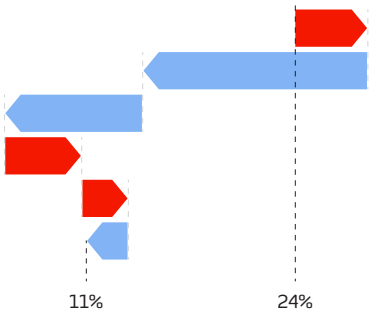


Figure 23: Illustration of how SHAP values take us from the cohort-level grand mean prediction (24%) to the individual prediction (11%). Arrows represent features (e.g. age and comorbidiy): red push the risk upward, blue push it downward.

[21] E.g. patients or admissions

II

Studies at a glance

The full manuscripts are included in part III, but this chapter outlines their rationales, study designs, key methodologies and main results.

Chapter contents

Prevalence and adverse outcomes of drug-drug interactions (I)

This study had a twofold purpose. First, to chart the landscape of potential drug-drug interactions (DDIs) prescribed at Danish hospitals and elicit patient types most prone to discouraged drug pairs. Second, to estimate the risk of adverse outcomes (length-of-stay, rehospitalisation and all-cause mortality) associated with discouraged drug pairs. Full manuscript on page 85.

Data

We used inpatient drug-prescription and register data between January 2008 and June 2016 from the BigTempHealth data set, including only admissions of individuals with concurrent use of at least two drugs. Successive in-hospital stays were combined into admissions if they were at most one day apart. As our reference for potential DDIs, we used Danish Drug Interactions Database (DID) [133]. This is maintained by the the Danish Medicines Agency (DMA) and covers mainly pharmacokinetic interactions.

Methods

In the first (descriptive) part of the study we used summary statistics and standardised differences in proportions (SPDs) (see page 45). The second (analytical) part used stratified Cox [64] and Poisson [134] regression models, with strata created by 1:5 matching patients on high-dimensional preference scores based on 843 features [135–138]. We fit outcome models only to patients whose preference scores were at least 0.3 but no greater than 0.7 [135].

Results

The full cohort comprised 2,886,227 admissions of 945,475 patients (54% female) with median age 62 years (inter-quartile range (IQR): 41-74) and a median drug load of 7 (IQR: 4-11). In the 1,836,170 admissions (of 659,525 patients) with ≥ 1 potential DDI, 54% of the patients were female, the median age was 65 years (IQR: 49-77), and the median drug load was 9 (IQR: 6-13).

The 18,192 patients (in 27,605 admissions) exposed to discouraged drug pairs were slightly older (median age: 68 years, IQR: 58-77), fewer were female (46%), and they had higher drug loads (median: 16, IQR: 11-22).

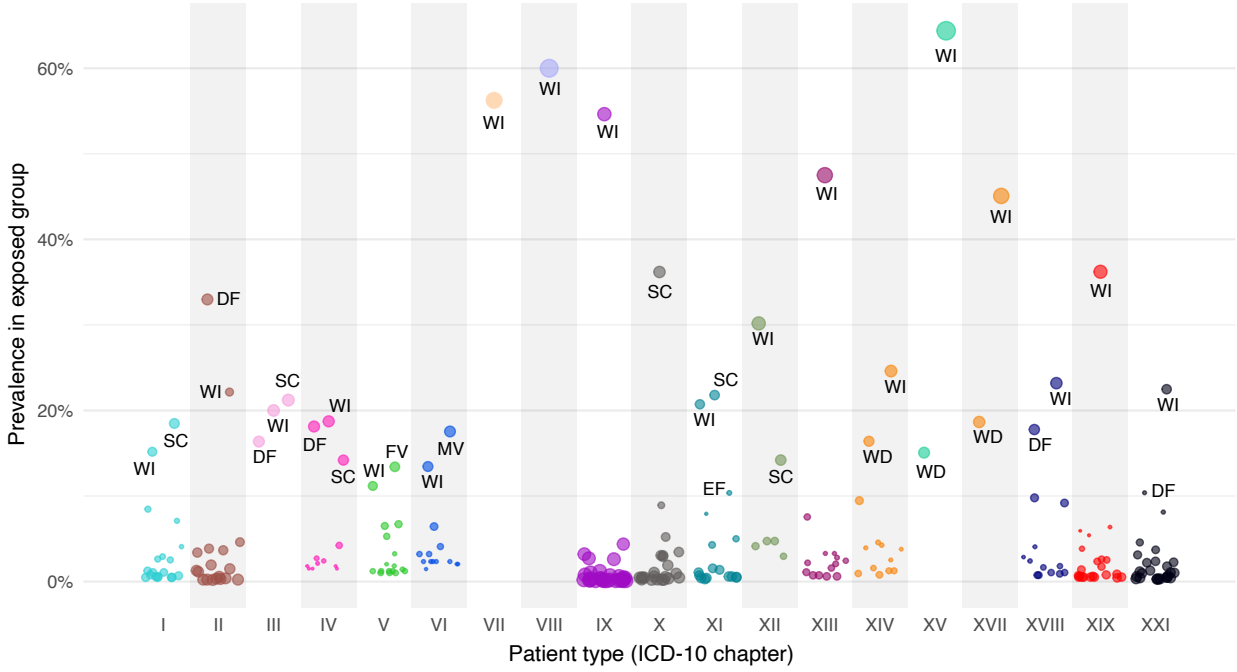


Figure 24: Prevalence of discouraged drug pairs by patient type. Each point represents one discouraged drug pair, and size the absolute value of the standardised difference in proportions using as reference admissions during which treatment with any discouraged pair was initiated. DF (N = 5): Domperidone (A03FA03) + Fluconazole (J02AC01); WD (N = 3): Warfarin (B01AA03) + Diclofenac (M01AB05, systemic); WI (N = 18): Warfarin (B01AA03) + Ibuprofen (M01AE01); SC (N = 6): Simvastatin (C10AA01) + Clarithromycin (J01FA09); MV (N = 1): Meropenem (J01DH02) + Valproic acid (N03AG01); EF (N = 1): Erythromycin (J01FA01) + Fluconazole (J02AC01); FV (N = 1): Fluoxetine (N06AB03) + Venlafaxine (N06AX16). Figure and caption reproduced as-is from figure 2 in study I.

In the 65 discouraged drug pairs (45%) prescribed to 5 patients or more, 7 were prevalently (>10% of admissions) prescribed during hospital admissions (figure 24). The most prominent pair was warfarin-ibuprofen, prevalent in all patient types except three (chapters X, XVI and XX). The second-most prominent was simvastatin-clarithromycin, prevalent in six patient types (I, III, IV and X-XII); the third-most was domperidone-fluconazole, prevalent in five patient types (II-IV, VXIII and XXI).

As figure 25 shows, the discouraged pairs meropenem-valproic acid, domperidone-fluconazole, imipramine-terbinafine, agomelatine-ciprofloxacin, clarithromycin-quetiapine and piroxicam-warfarin were associated with elevated mortality (shown with black points and lines in the figure). Confidence interval bounds of pairs associated with readmission were close to 1 and length-of-stay results were inconclusive.

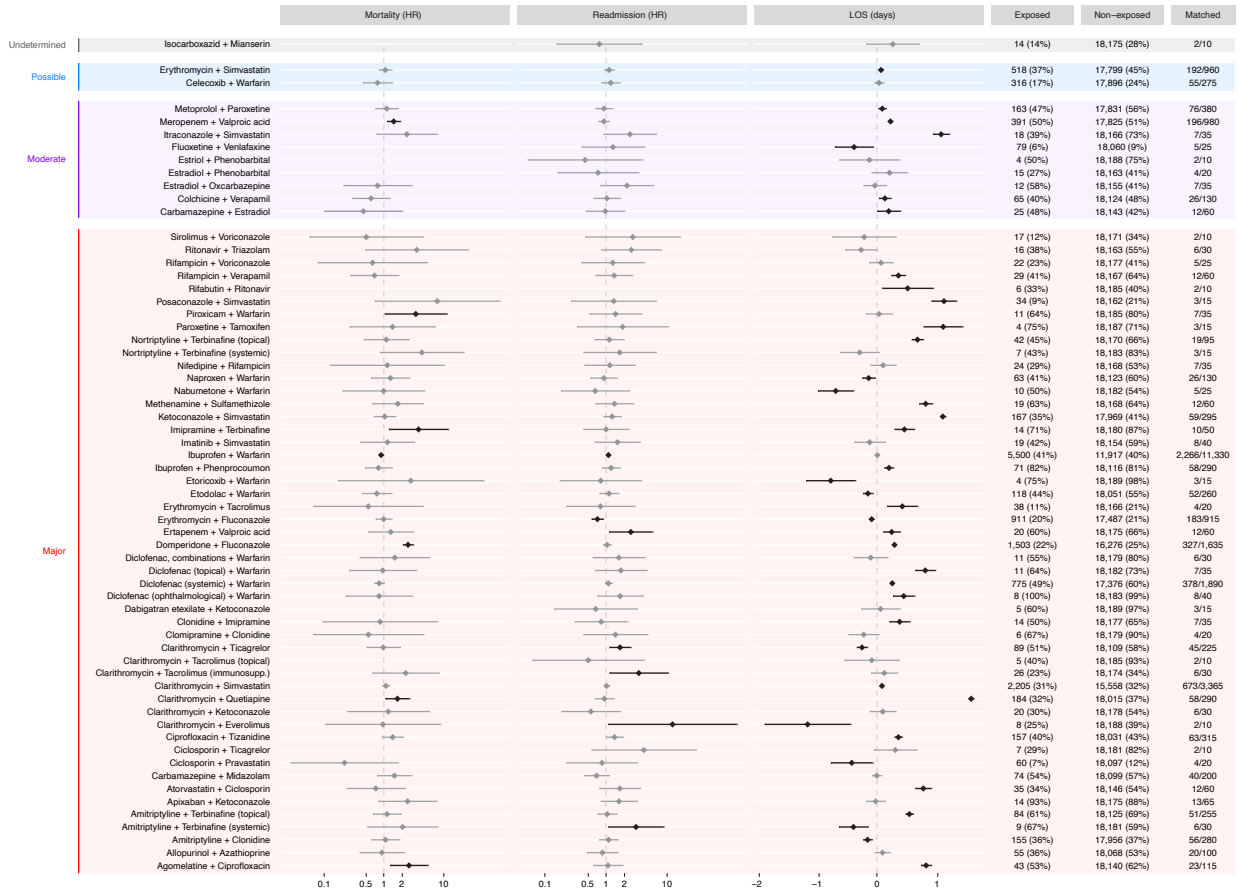


Figure 25: Estimate effect sizes of exposure to discouraged drug pairs and post-discharge mortality rate (hazard ratio, HR), readmission rate (HR) and length-of-stay (change in days). Diamonds show point estimates of the effect sizes, horisontal lines the 95% confidence intervals. The exposed and non-exposed columns show count (empirical equipoise) and the matched column shows the number of exposed/non-exposed used to estimate the effects of that pair. Figure and caption reproduced as-is from figure 3 in study I.

Conclusion

In this study, to our knowledge the largest of its kind, we found that well-known potential DDIs still abound, suggesting that pertinent information still goes unrecognised [139] and that point-of-care alerts may mitigate risk of harmful drug administrations. Our results suggest, in particular, that prescribing clinicians be alert when using strong inhibitor/inducer drugs (i.e. clarithromycin, valproic acid, terbinafine) and prevalent anticoagulants (i.e. warfarin and non-steroidal anti-inflammatory drugs) due to their great potential for harmful interactions. Our finding that 3A4 was the most prominent CYP isoenzyme involved in mortality and readmission rates agrees well with empirical evidence and clinical experience.

Renal dysfunction and inappropriate drug dosing (II)

In this study we sought to elicit the predictability of inappropriate drug dosing of select renal risk drugs to inform clinicians and healthcare personnel upfront about which patients with renal dysfunction are at elevated risk. We did so by crafting and comparing ridge logistic regression and multilayer perceptron (MLP) prediction models. Full manuscript on page 125.

Data

In this study we used in-patient drug-administration and biochemistry data along with the Danish National Patient Register (NPR) data, on admissions of adults between between 1 October 2009 and 1 June 2016. We set index at 24 hours after time of admission, including only admissions with at least one estimated glomerular filtration rate (eGFR) $\leq$ 30 mL/min/1.73 m^2 between time of admission and index.

Methods

We trained separate ridge logistic regression and MLP models to predict the risk of five outcomes: >0, ≥ 1, ≥ 2, ≥ 3 and ≥ 5 daily inappropriate doses. We crafted the hold-out test set with a time-series validation scheme. Hyperparameters were optimised with Optuna using 5-fold cross-validation, and prediction explained with SHapley Additive exPlanation (SHAP) values.

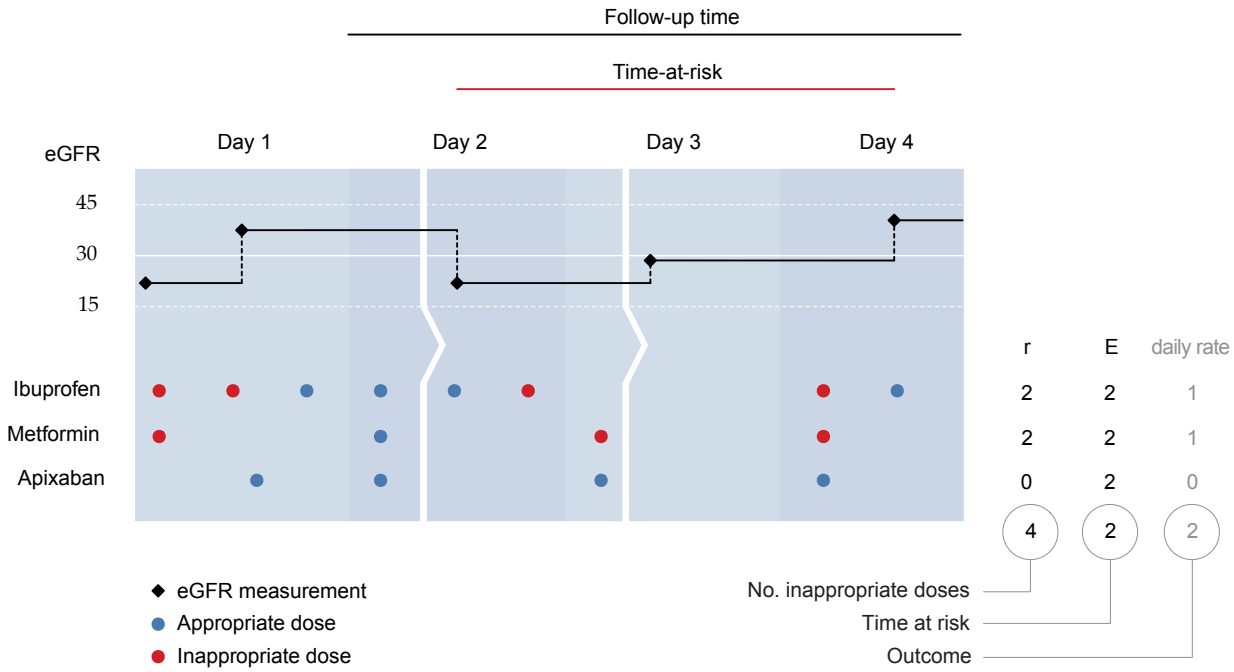


Figure 26: Deriving the outcome variables. This exemplary admission is composed of three successive in-patient visits (i.e. the patient has been transferred twice represented by the arrows). The admission is eligible because it spans more than 24 hours and an eGFR $\leq$ 30 was measured before index. Here, apixaban was given while the patient's eGFR was $\leq$ 30, but dose reduction rendered these administrations appropriate. Figure and caption reproduced as-is from figure 1 in study II.

Results

42,250 admissions (81%) started before 1 July 2015 and so we were used for model development, and the remaining 10,201 admissions made up the independent hold-out test set. The median age was 77 years and 50% of admissions were of women. $\geq$ 5 drugs were used between admission start and index in 23,124 admissions (44%). The most common drug classes used between admission and index were analgesics (Anatomic Therapeutic Chemical classification

(ATC) code N02, 37%), systemic antibacterials (J01, 35%), diuretics (C03, 33%), antithrombotics (B01, 28%), and antacids (A02, 25%).

The MLP models were slightly more performant, and were better calibrated in the development set, than their linear counterparts. All areas under the ROC curves (AUROCs) were between 0.77 and 0.81. Calibration in the test was about as good for ridge regression models as for MLP models.

Figure 27 illustrates how the MLP models picked up highly non-linear effects when such were appropriate. The SHAP values also gave rise to other insights. First and foremost, many features contribute substantively to the predictions of daily rate > 0 and ≥ 1 outcomes, while few features almost entirely drive the predictions for the other outcomes. Second, few features are the dominant prediction drivers across outcomes and models: use of anti-inflammatory, antirheumatic, and antidiabetic drugs as well as diagnoses of chronic kidney failure. Third, sex and age contribute little to predictions. Fourth, more pronounced polypharmacy pushes the risk up and vice-versa. Fifth, the linear models tend to give most weight to relatively few features whereas the MLP models spread out the contributions across more features. Finally, frailty (expressed as the number of previous admissions) was a more important driver for rarer outcomes, in the MLP models.

Conclusion

The trained predictions models can flag patients at high risk of receiving at least one inappropriate dose daily in a controlled in-silico setting. Using MLPs yielded some performance gains although slightly more involved ridge logistic regressions potentially might have been on par with these. A prospective clinical study would be needed to confirm (or refute) this in a real, clinical setting and whether this may translate into benefits in hard endpoints.

Language-agnostic safety signal detection in clinical notes (III)

In this study we combined several methodologies from machine learning, natural language processing (NLP), and data science to construct an end-to-end pipeline that produces safety signals for single-drug and drug-pair exposure. Its salient strength is that textual data need by neither in English nor curated. The latter comes about by turning things upside down: instead of predicting a likely outcome of a range of exposures, it yields likely exposures a given reaction, input as free text. Full manuscript on page 173.

Data

For this study we used all admissions of 500,000 randomly-sampled adult inpatients from the BigTempHealth cohort, making use of their in-hospital prescription data and free-text clinical notes

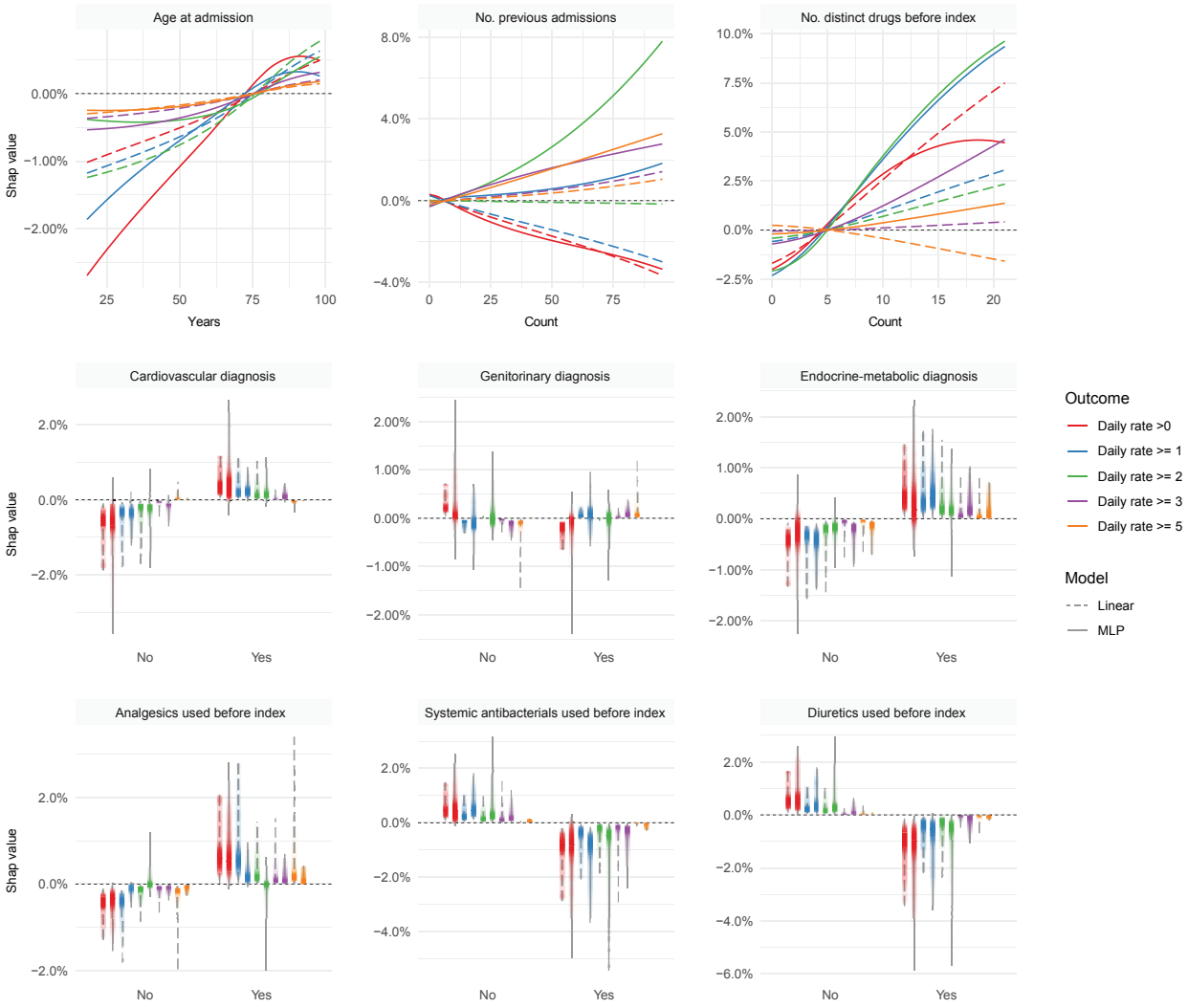


Figure 27: Bivariate relationships between values of select features (x axis) and their corresponding SHAP values (y axis). The continuous features are summarized by locally estimated scatterplot smoothing (LOESS), binary features by vertical density bands. Figure and caption reproduced as-is from study II.

recorded within 48 hours after time of admission. We only considered single drugs or drug pairs in at least 1,000 doorstep medication profiles.

Methods

Upon admission, the physician must synchronise the hospital medication system with the outpatient medication profile (we call this doorstep medication profile) before any changes are made, in order to properly document such changes.

To leverage the free text recorded in clinical notes, we used fastText [140] to train a 256-dimensional word embedding model from the full corpus, and terms from the included clinical notes that passed several filtering and processing steps: removing negated words, special characters, stop words, and Danish names. Importantly, we neither lemmatised nor stemmed words to retain natural words as the input. To associate each single-drug and drug-pair exposure with free-text information we used one MLP model with 2

hidden layers with each 256 nodes.

Results

The included admissions spanned the period 18 May 2008 through 30 June 2016 and comprised 2,905,251 admissions (54% of women), 10,788,259 clinical notes and 13,740,564 drug prescriptions. The final textual data contained 179,441,739 tokens for training the 10,720 MLP models of which 571 were for single drugs and the rest for drug pairs.

Figure 28: Fingerprint plots of 8 main UKU terms and their 571 single-drug signals. Inner circles: each wedge represents one drug and transparency the signal score. Outer circles: colours represent anatomical drug classes (ATC level 1). See caption of figures 2 (page 190) and 3 (page 191) in the full article for drug-class names and their colour coding. Subset of fingerprints from figure 3 in study III.

The fingerprints for 8 main UKU terms in figure 28 illustrate the single-drug exposures. These fingerprint plots illustrate that general or vague terms (bottom row) are relatively strongly associated with many drug exposures, and that for more specific terms (top row) fewer drugs of appropriate drug classes light up. Also, fingerprints of clinically related terms (e.g. tremor, parkinsonism and dystonia) are similar but clearly distinct from those of other terms (seen in figure 3 in full manuscript).

Safety signals (each independently assessed by two coauthors with moderate agreement as per Cohen's κ [141]) were generally meaningful, and terms with similar clinical meanings did yield similar exposure profiles. As figure 29 shows, 28 single-drug safety signals (8.1% of 345) were potentially undescribed or unknown; 16 drug-pair safety signals (14% of 115) were possible interactions.

Conclusion

Combining various flavours of machine learning and data scientific tools we successfully built an end-to-end pipeline for safety signal detection in medication and non-English textual data without the need for manual curation. We achieved this by turning

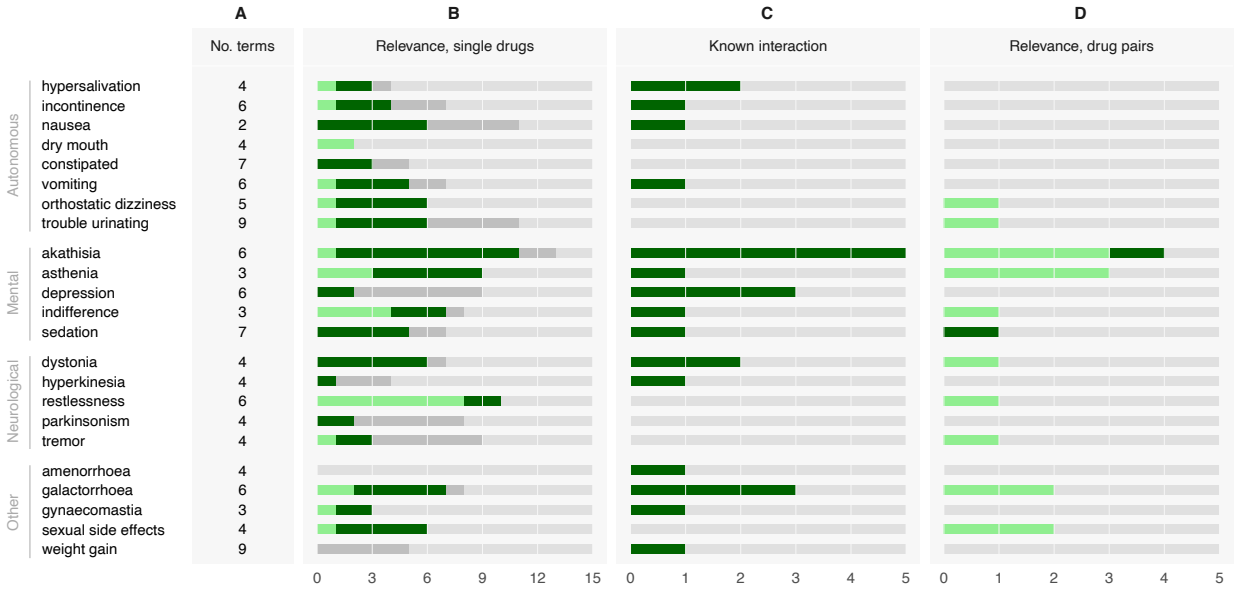


Figure 29: Main UKU terms by domain. Panel A shows the number of terms used in congruence analysis (total = 116). Panel B comprises all 345 single-drug assessments (23 terms x 5 single-drug signals = 115; 23 terms x 5 drug-pair signals x 2 drugs per pair = 230). Bright green: reaction possibly caused by single-drug (panel B) or drug-pair (panel D) exposure. Dark green: known reaction (panels B+D) or interaction (panel C). Dark grey: protopathic or indication bias. Light grey: spurious signal. Figure and caption reproduced as-is from figure 5 in study III.

things upside down, predicting not the likely outcome of a range of exposures, but the likely exposures for given outcomes of interest.

The congruence analysis suggests that the method pick up pertinent information, even when supplied with synonyms. With 8% of single-drug and 14% of drug-pair signals being possibly undescribed and relevant, its *hit rate* was in the expected order of magnitude [21, 142] and appropriate for its purpose: shortlisting few relevant safety signals from thousands of noisy ditto. These shortlists would then undergo review by pharmacologists, pharmacists, or other pharmacovigilantees and could complement existing safety signal detection in e.g. individual case safety reports (ICSRs).

Our approach is original in the field of adverse drug reaction safety signal detection and helps overcome many limitations of NLP methods relying on curated textual data in English. Crucially, this makes our method appealing in settings that must make sense of non-English free text for pharmacovigilance while, with few modifications, potentially lending itself well to alternative use cases such as patient-level decision-making support and drug repurposing.

Discussion

Leveraging longitudinal observational data, in the form of combined electronic health record (EHR) and electronic medical record (EMR) data, to answer pharmacovigilance questions proved much more difficult than anticipated. This chapter discusses key methodological challenges, weaknesses, and reflections pertaining to our studies, through the lens of the three questions posed in the Scope.

Chapter contents

Outcome operationalisation

Undertaking three separate studies, we had to overcome several challenges and make some decisions about operationalising outcomes.

Study I

We used standardised differences in proportions (SPDs) to highlight drug exposures more prevalent in certain patient types than others (see also figure 2 of the manuscript, page 104). Using a reporting odds ratio had likely left the results qualitatively unaltered, but a viable alternative—and in hindsight conceptually appealing—method had been the shrinkage log odds ratio (SLOR) from disproportionality analysis of safety signals [143]. Because it is shrunken, it might better account for rare drug pairs, and its Bayesian interpretation might have provided better statistical properties than that of SPD for ranking by imbalance.

Study II

In study II we originally intended to use hurdle models with an offset Poisson component to model the daily rates of inappropriate doses. A hurdle model has two components [144]: a binary regression submodel for whether the patient received any inappropriate doses, and a count submodel for the expected daily rate of inappropriate doses using e.g. a truncated Poisson distribution. As such, the hurdle model can be considered a discrete mixture model in which one process governs the zero counts and another the positive, non-zero counts [145, sec. 4.3].

We had built the architecture for such a model, but for two principal reasons we decided to go with the final approach of dichotomising the outcomes and training several binary-regression models instead. First, the hurdle-model setup would impose a parametric outcome model (in the second component) that might fit the data-generating process poorly or even hamper learning due to more complex cost functions. The binary-only outcome approach imposes fewer assumptions on the modelling and eases evaluation; also, this way the multilayer perceptron (MLP) models remain non-parametric.

Second, the purpose of developing such a prediction model was, anyway, to flag patients at elevated risk of receiving inappropriate doses. So, even with a hurdle model (or any modelling the daily rate of inappropriate doses as a numeric outcome) we would need to set a threshold somewhere. By dichotomising the outcome upfront, this decision is moved upstream and aids resolving the potential problems with parametric outcome models.

However, even with a dichotomised outcome we might be interested in estimating the uncertainty about predicted risks. MLPs do not lend themselves well to fully-Bayesian analysis because of their complex posterior distributions (causing convergence problems), but modelling the risks with a Beta distribution (with two parameters) instead of a Bernoulli (with one parameter) could be a viable approach, and one rarely seen for binary regression in machine learning.

Another limitation of the outcome variable *daily rate of inappropriate doses* (on which the dichotomised predicted outcomes were based) is that it is a soft endpoint. We do not actually know if it is a good proxy for noxious outcomes or not: even if patients are exposed to higher doses of drugs than they should be, they may remain unaffected. For example, it is clinically plausible that a patient suffers no harm even if they continue their metformin treatment for a few days despite dehydration-induced transient kidney dysfunction.

One way to include noxious outcomes—such as prolonged hospitalisation, readmission, and mortality—could have been using a Hidden Markov Model (HMM) setup [146], modelling different states from index to end of follow-up. Another might be to employ

causal discovery, perhaps utilising the temporal information in our data with the temporal PC algorithm [147] although this was unavailable when the study was conducted.

Study III

We took a fairly simplistic approach when defining the drug-pair outcomes in this study: a patient was considered exposed to a given drug pair or not. However, the patient could also be exposed to just one of the drugs, and including that information in the model might alter (and perhaps improve the quality of) the safety signals.

Thus, a factorial design could have been a better way to model drug-pair safety signals, for example using 4 mutually exclusive outcome nodes in each model (exposure to none of the drugs, exposure to drug 1 only, exposure to drug 2 only, exposure to both drugs) with a `softmax` activation function and a `categorical crossentropy` loss function.

Explained predictions ≠ causal relationships

There is growing awareness that machine learning prediction models do not yield causal relationships by themselves [148]; indeed, notwithstanding sophisticated attempts at explaining predictions, such models will perpetuate (and perhaps corroborate) biases captured in the development data [79], a clear sign that they fail to recognise causal relationships and only capture associational ditto. Therefore, when the purpose is outcome prediction, frameworks that yield prediction drivers become useful for *sanity checks* to ascertain, to the extent possible, that the model picked up pertinent information in the data, and not biases or misleading proxies. Using i.a. SHapley Additive exPlanation (SHAP) values this way (as we did in study II) helps improve prediction models.

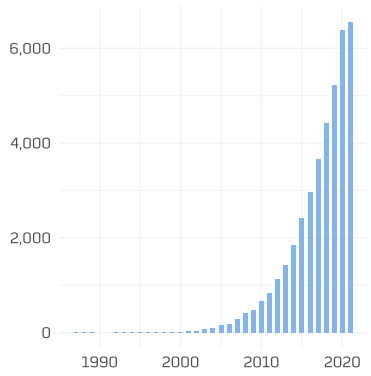


Figure 30: Number of hits on PubMed (y axis) by year (x axis). Query on 21 October 2021: `"propensity score*"`

Causal inference in observational data is possible; methods based on propensity scores enjoy increasing usage (figure 30) and are perhaps the most widespread approach. Propensity scores can be used in a variety of ways, including weighted regression, regression adjustment, matching, and stratification [149, ch. 12] These are sometimes referred to as pseudo-randomisation because they seek to accomplish what randomised controlled trials (RCTs) do: apart from random error, the difference in outcome between groups is attributable to the exposure under study [150].

The validity of such methods rests on several assumptions: a well-defined equivalent trial exists, absence of unmeasured confounding,[22] consistency, positivity, and no interference [149]. Even very large data sets such as that of BigTempHealth cannot guarantee unmeasured confounding: much information escapes clinical and administrative data, perhaps most notably socio-economic status and changing clinical practice.

Using preference scores[23] in study I (page 91) we did not seek to

[22] Related to epistemic uncertainty [151]

[23] A slight adaptation of raw propensity scores [135]

make definitive claims of causality but instead use a well grounded method for finding interesting associations with stronger real-life bearing that blindly hammering variables into a generalised linear model (GLM) and hope for the best, or (perhaps worse) blindly using a machine learning model with SHAP values.

Equipoise is "a state of genuine uncertainty about the benefits or harms that may result from different exposures or interventions" [67, p. 84] and a key notion in RCTs. Equivalently, in causal inference based on propensity scores, *empirical equipoise*[24] is an important, if somewhat overlooked, consideration. Mindful of this, the percentages in the *Exposed* and *Non-exposed* columns in figure 25 show that empirical equipoise could only reasonably be assumed in 22 exposure groups, e.g. *Meropenem + Valproic acid*, *Paroxetine + Tamoxifen* and *Clonidine + Imipramine*.

[24] "Accept drug pairs as emerging from empirical equipoise if at least half of the dispensings of each of the drugs are to patients with a preference score of between 0.3 and 0.7." [135, p. 12]

What is perhaps worse, the final matched cohorts, in which the Poisson and Cox regression models were fit, were generally very small for exposures with empirical equipoise. This naturally calls into question the generalisability of the findings and warrants further investigation (e.g. with discrepancies in summary statistics of those matched and the full cohort). Thus, as was also highlighted by Walker et al. [135, p. 19], equivalent models should be fit in different populations, ideally using heterogeneous data sources, to better understand the wider applicability of the results. Exactly the way so-called Observational Health Data Sciences and Informatics program (OHDSI) network studies are carried out.

Our approach to modelling the exposure in study I was relatively simplistic, in that we considered exposure to the drug pair a binary thing. A more appropriate approach had probably been regressing the adverse outcomes (length-of-stay, readmission, and all-cause post-discharge mortality) on independent effects of each drug in the drug pair as well as an interaction term. Even though that would add a few more parameters in the model, one could argue that the results (considering the reasonably large number of admissions from which they arose) would benefit enough to warrant this approach.

In our setup, the high-dimensional lasso logistic regression models served as prediction models to yield one propensity score for each patient. Proper performance evaluation (as per Evaluating prediction models) had likely been warranted to ascertain well-calibrated models: poor calibration might be problematic because we matched patients on their preference scores, derived from the predicted propensity scores.

A conceptually and methodologically different approach would have been to use causal discovery [152, 153] that, in some sense, bridges the gap between machine learning and causal inference. Recent advances in this field leverage temporal information [147], but these were unavailable when the study was conducted.

It would have been interesting to see if results similar to ours would emerge in other data sets arising from healthcare in other

cultural and health-political contexts. Although we did use tools from the OHDSI ecosystem, our data model was not compatible with the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) and so did not allow for simple comparison with data sites in the network.

Secondary data

Whether data be primary or secondary is less important than having data appropriate for our research questions [64, ch. 23]. To be sure, using secondary data restricts the spectrum of questions we can and should ask; a very recent and sobering example is the surge in COVID-19 studies hinging on secondary EHR data, with unfortunate consequences when their inherent characteristics went unrecognised [154].

The salient point is that the primary purpose of our secondary data is caring for patients and managing the healthcare system. So, at their core these data must provide *clinical utility* while we seek to unleash their *scientific utility*. For example, data designed for process support[25] are not necessarily ideal for research. Thus, tapping into such data poses multiple problems starting with something as basic as how to define an *admission*: if departments are seen as distinct entities, considering stays at each department contiguous admissions is sensible, but from a scientific perspective this fragmentation of the full in-hospital visits must be reconciled or perhaps used actively, e.g. if the interest revolves around risk of intensive care unit (ICU) transfer.

[25] Such as how the patient moves around in the healthcare system as per the Danish Contact Model [155, pp. 13–14]

Another example is that of lingering Electronic Patient Medication 1 (EPM1) records in the Electronic Patient Medication 3 (EPM3) data set. This was a complication we needed to handle in the extract-transform-load (ETL) process, but this may make perfect sense from a clinical vantage point: healthcare staff needed to have the medical profiles of patients who visited the hospital in the early period after transitioning from EPM1 to EPM3, and a simple way to do that could have been to simply synchronise the two systems.

Nonetheless, when used correctly longitudinal observational data arising from EHRs can fill in the evidence gaps left by RCTs [156] and for economic evaluations [157, ch. 8]. Sometimes, even, "secondary data might be the best source given the available resources" [64, p. 481] and the ability to follow patients over (long) periods of time is an important tool in safety profiling of medicines as, only then, may outcomes of long-term exposure or late-onset effects surface [158]. Thus, having large-scale systems with curated and operational EHR data in place allows researchers to undertake the kind of sound clinical research essential for society to make well-grounded decisions [159].

Despite great efforts and successes in making clinical systems communicate, inter-sector data transfer often passes through central databases, such as the so-called cloud-based *shared medication record*

(fælles medicinkort, FMK) [160]. Whenever a patient is admitted to a hospital, the current medication profile is downloaded to the hospital's medication module and in-hospital modifications occur in this local copy. At discharge, the local copy of the medication profile is, then, uploaded to the shared medication record for other practitioners to see and modify. This system works well and underpins clinical work as long as each electronic patient system and the shared medication record platform communicate and do so well (unfortunately not always the case).

The very same system, however, causes a great deal of trouble in subsequent analyses, as hospital records offer little information about genuine medication use between hospital visits. In the ideal world in which all medication data (community-pharmacy and in-hospital dispensations alike) were kept in one place, such as the shared medication card, we could craft better exposure profiles that might elicit safety problems with medicine use. So, this kind of data siloing is yet another thing that necessitates reconciling drug-safety data when using secondary data for pharmacovigilance.

Error-prone preprocessing

Valid data analysis hinges on sound data preprocessing that does not distort the data. As described in Data, our raw biochemistry and medication data received left much to be desired, especially from a documentation viewpoint. This meant that many resources went into deciphering and disentangling the data to cast them into an operational format suitable for the BigTempHealth research goals, including our studies in the realms of pharmacovigilance.

Undertaking such full-scale ETL processes with little or inadequate documentation is errorprone. In hindsight, a validation study akin to what others have done for Electronic Patient Medication (EPM) [89], but not restricted to medication, might have been a great resource for pinpointing problems arising from erroneous preprocessing.

More than a time-consuming annoyance without direct scientific output, it is wasteful that the same data are processed and operationalised over and over again by different groups and stakeholders, a notion also raised by others [161, 162]. It also complicates reproducibility in science because the results of any statistical analysis will depend on the data input: if preprocessing steps of the same raw data differ, so will the results.

Myriad pecularities needed resolving, and here I present three illustrative challenges relevant to our studies. First, there are different ways to record *how much* of a drug was given to a patient. The EPM data needed little work as they directly gave the dose and its unit: if 2 tablets of 500 mg. metformin were administered, it was recorded as an administration of 1000 mg. metformin. Such data simplifies validation as we, essentially, need only confirm that doses and units be meaningful together, and that doses be reasonable.

OPUS-medicin had a different approach: for every administration, we had to compute the total given dose by combining the strength and its unit with the given amount and its unit. So, we would have the information that 2 tablets of metformin with a strength of 500 mg./tablet were given. When the units line up, the total dose is just the product of these two numbers and some simple logic to deduce the unit of the total dose.[26] Sometimes they do not, and we had to handle cases like 2 drops of metformin with a strength of 500 mg./tablet. This quickly becomes unwieldy due to the huge number of potential combinations and is caused by insufficient (if existent) data-entry validation in the medication module. Curiously, we also encountered not so few cases in which the doses more than anything else resembled Danish phone numbers. For the purpose of study II, the only study in which we used the actual doses administered, manually checking inconsistencies was time-consuming but possible because we only considered 7 drugs. But scaling this effort to cover substantial parts of the full data set would be impractical.

[26] For example: x tablets $\times$ y mg./tablet $= z$ mg.

Second, comprehensive preprocessing was needed for the B-Data Clinical Chemistry Laboratory System (BCC) data. It turned out that Nomenclature, Properties and Units (NPU) codes were inconsistent and not necessarily uniquely linked to specific components (e.g. potassium and sodium) in specific specimens (e.g. blood or urine). It makes a large clinical difference whether glucose is measured in blood or urine, and so within BigTempHealth an alternative (largely manually created) coding scheme was used, see Idiosyncratic data modelling.

Third, the issue of possibly duplicate clinical notes emerged. Their origin was undocumented but seemed related to our having several iterations of the same clinical notes (perhaps drafts saved while collecting more data and observations). Thus, much of the text would be identical but there would be differences, and only the final version would be useful for our needs. Setting up automated de-duplication without erroneously removing legitimate clinical notes proved remarkably difficult. Related to this are *smart phrases* that take form-like structure but nonetheless store everything as free text. Such standardised text can also make deduplication difficult as they may be very alike but genuinely pertain to different patients.

Only in-hospital data

Part of the BigTempHealth research programme, we had access only to in-hospital data which restricted the types of questions we could pose and, consequently, the studies we could undertake. This limitation has four main components. First, albeit comprehensive with respect to both population and longevity, the the Danish National Patient Register (NPR) has a key weakness: it holds no primary-sector information on diagnoses, procedures, or treatments.

Although we did not have access to them, Danish out-of-hospital healthcare utilisation data exist [83, 163] but do not hold detailed phenotypic patient data (such as diagnoses) as it is outside the scope of the register.

Second, we have no real records of out-of-hospital medication exposure. Instead, we resorted to what we called *doorstep medication profiles* in study III (see 179). Although this is probably a decent proxy, having the actual data from the Drug Statistics Register [83, 164] had likely yielded more accurate exposure profiles at time of admission. Of note, the consumption of over-the-counter drugs and poor compliance with prescription drugs escape both the Drug Statistics Register and our doorstep medication profiles although the former hold some data on over-the-counter drugs.

Third, the BigTempHealth cohort only comprised patients who visited the hospital and did so between 1 January 2006 and 1 July 2016. This selection (and consequent potential for bias) must be borne in mind, and the results of our analyses seen in light of this: the average patient seen at a hospital is likely sicker and older than an average patient seen by general practitioners or, indeed, the average person in the population.

Fourth, for many patients in the BigTempHealth cohort, our data on all axes are left-censored and for some (such as mortality and readmission) also right-censored. This hinders many types of studies of the relationship between exposure and outcome as we often have no well-defined start of exposure. We can use a *new user*-design and follow patients who start a new treatment during their admission, but this is not entirely unproblematic: they may not be truly new users, and if we condition on patients having previous admissions without exposure, we also condition on patients probably being worse off than the background population (as also touched on above).

Idiosyncratic data modelling

Obtaining data can be difficult enough, but the work only then really begins: big data are not necessarily good data, and raw data are usually unfit for analysis without preprocessing involving setting up ETL processes to yield operational data in some data model [165] that can be stored somehow, whether as flat files or in a database. This becomes problematic when each research group must (or at least does) waste valuable resources undertaking their own variations of data preprocessing with little or no insights into exactly *what* was done [161].

Sygehusvæsenets Klassifikationssystem (SKS) is the data model of NPR and somewhat specific to Denmark.[27] Such specific data models make it difficult to undertake large-scale studies with data from other countries. In particular, analytic code will necessarily be specific to the data model for which it was developed. Thus, the code we have produced (and, for studies II and III, shared to

[27] Diagnoses are encoded with a Danish dialect of the the 10th revision of the International Classification of Disease (ICD10) and procedures with the NOMESCO Classification of Surgical Procedures (NCSP)

ensure transparency and aid scrutiny [166, 167]) will need adaption to become appropriate for other researchers, even if their data originate from the same sources.

The issue of idiosyncratic data models was particularly acute for our biochemistry data. Even though both data sources seemed to use codes from a standardised nomenclature (NPU) with links to International Union of Pure and Applied Chemistry (IUPAC) codes, it turned out that regional variations in these codes rendered them near-useless for identifying specific biochemical analyses. Instead, we ended up building in-house standards for components (e.g. potassium and glucose) and specimens (e.g. plasma and urine). This enabled querying with with post-hoc reconciliation of units (if needed), somewhat akin to post-coordination [168]. We had to do this in a study we prepared [169] but decided to defer in the interest of time and resources as it fell outside the scope of the project.

Textual data

Free text is a wonderful medium for data collection and presentation because it gives the author full control over what is recorded and how it is presented, as discussed previously (page 34).

The textual data in clinical notes proved amazingly untidy with abounding ambiguities attributable to different "dialects" between medical specialties: the meaning of *GA*, for example, would likely be gestational age in obstetrics but general anaethesia in surgical specialties (including gynaecology). Ambiguities and the use of abbrevations is probably also somewhat correlated with seniority and experience level, and perhaps also the setting in which the note was written (e.g. inpatients vs. outpatients).

There seemed no good way to overcome this fundamental characteristic of the data when used the way we did in study III. We trained the embedding model on our own data, but it might be reasonable to expect that an existing language model, pre-trained on a relevant corpus in Danish, could yield better results. The problem with this notion is that there are not many relevant corpora available. Even training such a model on medical textbooks in Danish might be suboptimal because the language used therein is proper Danish in stark contrast to clinical notes.

Thus, despite the elegance in how embedding models operationalise textual, for brute-force approaches such as that in study III, simpler methods building on (maybe fuzzy) matching might be more viable while also methodologically simpler.

Complex analyses

In 2000 Edwards and Aronson stated that that cohort and case-control studies for pharmacovigilance requires complex calculations [14]. The need for complex calculations has only gone one way

since then and now pharmacovigilantes can reap the benefits of the tremendous advances in data science, statisics, and engineering to leverage huge data sets.

Scientific articles cannot generally accommodate all technical details of their underlying analyses, and so referees and readers must consult the analytic code directly to truly understand how the results came about. The codebases of our three studies make up thousands of lines of code in various languages (mainly R, Python, and SQL) and while we have done what we could to ensure that our code be correct and reflect what we report in the manuscripts, there is always the risk of errors.

Scientific software is neither meant nor designed for production and often leaves much to be desired from a programming point of view, is prone to errors, and is difficult to scrutinise [166, 167] even for skilled peers. We cannot always share our data, especially secondary data as ours. Thus, even if analytic code is publicly available—as we did for studies II and III—it is very difficult to scrutinise, or even just comprehend, such code if it were not crafted for a common data model.

Moreover, even the two end-to-end analytic pipelines we shared (study II and study III) do not truly start at the beginning. As already discussed, when raw data arrive from private vendors or public registers, they undergo preprocessing and these steps (perhaps even more errorprone than the processing in the analytic pipeline) are usually not shared anywhere.[28]

[28] In the OHDSI community, sharing the ETL code is actually encouraged so others can learn from previous experience

Some have called for researchers to share their data formatting keys [161], but this misses a salient charateristic of modern epidemiology: we no longer use only tidy structured data whose encoding need reconciling. Rather, many applications in analytical epidemiology using machine learning leverage e.g. waveform and textual data that require a radically different approaches. Thus, we need to share not merely formatting keys but full analytic workflows. Snakemake [170], born in bioinformatics and used in all three studies, is a powerful framework for building analytic pipelines, but alternatives such as Nextflow [171] do exist.

These frameworks encourage modularisation: breaking down code into manageable and digestible chunks. Snakemake and the like are still uncommon in epidemiology, but having a unified language for building code along with the adoption of common data models would likely remedy many of the issues discussed in this section. One issue that cannot be resolved, however, is that of the computational requirements of these pipelines.

Very complex (machine learning) analytic pipelines tend to require specialised hardware to run and finish within a reasonable timeframe [106]. Study III is a good example: it took more than 24 hours to train the embedding model even though it had 120 central processing unit (CPU) cores and 200 GB memory at its disposal.[29]

[29] As a comparison, my computer has 4 CPU cores and 16 GB memory

Over-engineered solutions

Machine learning covers many disciplines and machine learning models can take many forms. To focus our efforts, we used only MLP models throughout and so perhaps fell victim to the *law of the instrument*: if all you have is a hammer, everything looks like a nail; MLPs became our hammer.

In study II, however, simpler model types such as support vector machines or random-forest models could have been used in place of the MLPs, might have been as performant, and had likely trained faster.

In a similar vein, simpler but equally performant alternatives in study III might have been viable. For example, one could take an input term—representing a potential adverse drug reaction (ADR)—and use the learnt embedding model to find, say, *clinical cousins*[30] similar to some degree, e.g. top-5 or with cosine similarity > 0.9. Then, one could query the retained tokens for the presence of any of these derived target tokens, to construct a 2-by-2 exposure-outcome contingency table and use the SLOR, akin to the way it is used in disproportionality analysis in individual case safety reports (ICSRs). This might even handle the imbalanced nature of the problem better than our randomly under-sampling the majority class, a classic tack in machine learning.

[30] In study III we considered *clinical cousins* terms that mean (almost) the same thing but may be lexicographically very different

Such an approach would definitely be faster in the development phase and it would render pertinence evaluation superflous. It would, however, require full access to the entire corpus at query time. Storing almost 11 million clinical notes in a database with several million doorstep medication profiles[31] may not be feasible. On the other hand, keeping some 4,000 MLP binary files (each about 2 MB) is perhaps not more desirable in the end, from a resource-requirement perspective.

[31] As well as supporting data structures such as indices for tolerable querying time

An additional (major) benefit of the contingency-table approach would be that another disproportionality method, the Ω shrinkage measure [172], would be available for (more) proper statistical modelling of drug-drug interactions (DDIs) in this kind of safety signal analysis.

Conclusion and outlook

With the increasing computerisation of healthcare, and subsequent gathering of massive amounts of observational longitudinal data, the hope was that these data sources would easily lend themselves well to pharmacovigilance and perhaps even supersede individual case safety reports (ICSRs). The former proved much more difficult than anticipated, and based on the following discussion of the research questions—in turn building on our experiences and the reflections they sparked—I conjecture that we can utilise longitudinal observational data in two ways to viably complement existing spontaneous reporting systems (SRSs), not replace them.

First, by using existing, or partaking in building new data warehouses with data models appropriate for pharmacovigilance. This notion feeds into the idea of *the learning healthcare system* [173]. Although there might be other viable options, from our experience Observational Medical Outcomes Partnership (OMOP) is one prudent choice of common data model (CDM): it was born and is developed with pharmacoepidemiology and pharmacovigilance in mind at the nexus between regulatory bodies, acadaemia, and the industry [165, 174].

Second, by exploiting frameworks based on CDMs and federated privacy-by-design [175–177] data analysis with post-hoc aggregation of results, not data. Other than addressing privacy issues, this delinks the ability to generate original ideas and intelligent questions from that of data access: even though you have never seen their data and never will, you know the structure of the data of everybody else and can nevertheless design studies and build the required analytic code, which is then shared across the data network. This will bring more minds and ideas to the table, ultimately improving the questions asked and studies conducted to answer them.

Chapter contents

Can observational data cater for pharmacovigilance?

The short answer to the question is, "it depends". The nature of the BigTempHealth cohort turned out to put some hefty restrictions on the kinds of questions we could reasonably ask. In particular, the very limited medication data precluded well-defined exposure operationalisation for most kinds of typical questions in pharmacovigilance and pharmacoepidemiology in general.

Having access to very detailed in-hospital data did, however, enable us to undertake studies beyond what it possible with register data only. In particular, the operationalisation of the outcomes in study II and the ability to tap into textual data from clinical notes in study III were novel and could inspire future research with these data types and sources.

Unlike registers, electronic medical record (EMR) offer a high level of granularity and detail, but this is not enough to be useful in pharmacovigilance: without data that underpin well-defined exposure operationalisation—a key component of pharmacovigilance—we cannot fully reap the benefits of high granularity in genuinely interesting research endeavours. I illustrate some examples in figure 31:

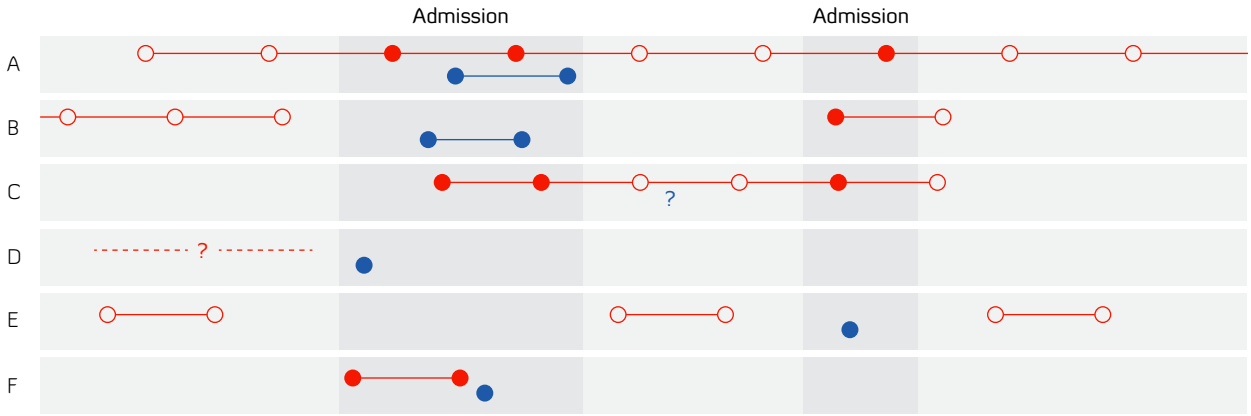


Figure 31: With in-hospital data many study designs become impractical due to ill-defined start and end of exposure. Each row represents one patient. Filled circles: observed. Hollow circles: unobserved, inferred from prescription records. Red: exposure to drug. Blue: outcome of interest.

Patients A and F represent the kind of questions we were best able to address: we directly observe the exposure and immediate adverse drug reactions (ADRs). Patients B and E are less desirable, but if we assume that the start-of-admission records be correct, we can infer exposure or lack thereof before the outcome. Patients C and D are worst: we have no way to gauge if the patient was exposed, and we cannot observe events that happen outside the hospital.

Obtaining complete medication data is difficult, if not impossible. In Denmark, however, the Drug Statistics Register [164] provides detailed information on purchased prescription drugs. Access to this data source might be necessary for observational data to truly cater for pharmacovigilance, in a Danish context.

What are the hindrances?

Here, I highlight 9 principal hindrances in 3 domains[32] that should be addressed upfront when considering using observational data for pharmacovigilance:

[32] Inspired by Johnson et al. [178]

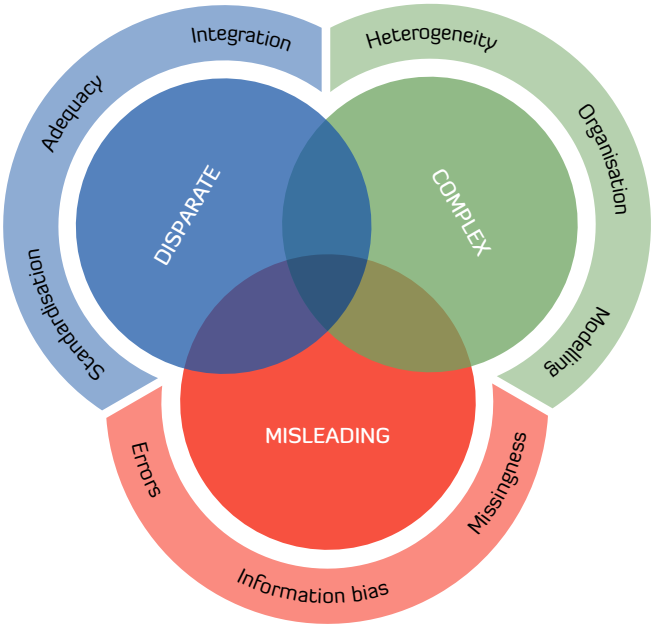


Figure 32: Hindrances to successfully using observational data in pharmacovigilance.

Complex data

Modelling Statistical modelling offers many ways to answer the abounding pharmacovigilance questions. We used methods from descriptive, aetiological, and predictive epidemiology (studies I and II) as well as safety signal detection (study III). A natural place to begin is to understand what kinds of questions one wants to pose, what kinds of modelling one intends to use to answer these and, crucially, if these are underpinned by the data.

Causal inference in observational data, for example, hinges on the assumption of no unmeasured confounding, so one should ensure access to data that will reasonably satisfy this assumption. Another example is safety signal detection that tends to revolve around rare patterns [143, p. 1259], which in turn requires large data sets. Although somewhat tangential to pharmacovigilance, drug repositioning can build on the same methods as safety signal detection, and so also in this regard having enough data is important [179].

Organisation It takes a village to exploit electronic health record (EHR) data for pharmacovigilance; here I consider organisation in a broad sense, to include also infrastructure. The path from raw data in siloes to operational data suitable for pharmacovigilant machine learning (as well as other applications) requires a diverse skillset (see figure 33) and a concerted effort. Such diversity can

only realistically be achived by an organisation that also has access to adequate hardware and IT infrastructure.

Heterogeneity Observational data can be heterogeneous, and have only become more so in the past decade. We used tabular, not-so-high-dimensional data and textual data, but other data types are available and becoming increasingly so, not least various kinds of omics data.

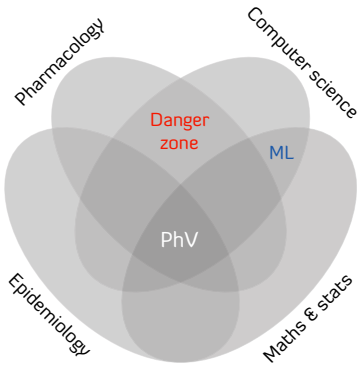


Figure 33: Venn diagram of pharmacovigilance (PhV). Inspired by Conway [180] and Alarcón-Soto et al. [181].

Disparate data

Integration Data collection is delinked from the analytic needs, so we answer our questions to the best of abilities with the data we can get our hands on. Observational data live in siloes and an important step is to be able to marshal them so they can be analysed together. This might become easier in the future with learning healthcare systems [173],[33] but it will likely remain a major hurdle that must be overcome.

[33] Akin to Facebook or Google whose data collection are intimately linked with the purpose of their analyses, enabling optimisation of the former to underpin the latter

Adequacy We need enough data, with respect to both quality and quantity. It may be unreaslistic to have enough data e.g. if the interest revolves around rare exposures and/or outcomes, if the geographic area giving rise to the data is too restricted, or if we do not cover a sufficiently long time span.

This is important for both aetiological, predictive, and safety-signal detection endeavours albeit in different ways. Statistical power is a key component in aetiological epidemiology, whereas too few observed events will preclude training any useful prediction model or sound validation. For safety signal detection, we need enough exposure-event pairs for them to stand out.

Standardisation Collating structured data from disparate sources may entail standardising and reconciling encoding to ensure that, say, a given diagnosis can be identified uniquely in all data sets. Employing a CDM will further facilitate collaboration with other data holders using the same CDM [182]. The Observational Health Data Sciences and Informatics program (OHDSI) network studies build on the notion that interesting research questions may come about in settings without data, but data may abound elsewhere. Because we know what the data look like in other sites, we can design statistical analyses to leverage these data without ever seeing them.

Misleading data

The subtitle for this domain might as well be the well-known *garbage in, garbage out*: the results of any analysis hinge entirely on good data. This applies no less to pharmacovigilance [182] and draws on the virtues of i.a. epidemiological prudence with respect to quality assessment of data.

Errors Errors can arise before we receive the data (at data collection) and after (during preprocessing). For example, we received several versions of the Electronic Patient Medication (EPM) data before we felt convinced that they were complete; even then, as described earlier, duplicate records lingered in the years around changing systems. One particularly tricky preprocessing challenge we faced was that of de-duplication of clinical notes. It proved difficult to set up an automated way to carry out this task without inadvertently discarding legitimate notes.

Information bias This comprises measurement error for continuous variables and misclassification for categorial ones [64, ch. 9].

Missingness This encompasses missing patients (i.e. selection bias) and missing data points for patients in the data set. Selection bias can happen due to e.g. Berksonian bias when pooling data sources [64, ch. 12]. Data missingness falls into three groups, of which the last is the most challenging: missing-completely-at-random (MCAR), missing-at-random (MAR) and missing-not-at-random (MNAR) [183].

Is it worth our while?

All three studies yielded interesting findings, showcasing that observational data such as those in the BigTempHealth research programme can underpin pharmacovigilance. Leveraging distributed data networks, e.g. by means of a CDM, would only amplify this potential.

We did not undertake any formal cost-effectiveness evaluation but collating, operationalising, maintaining, and analysing observational data are costly operations. Data collation is bureaucratic and appropriate staff to oversee this is key. As addressed repeatedly in this thesis, operationalising data is a large undertaking requiring people with diverse backgrounds and areas of expertise; this component also involves adequate hardware. If data collation is a one-off event, the maintenance amounts to costs pertaining to hardware so the data remain operational and available for analysis; with continual data capture the costs are greater and likely more unpredictable. Analysing normal epidemiological data can happen on a desktop computer and runtime can usually be counted in seconds or minutes. Machine learning in big observational data is another beast: running the analyses of studies II and III each took several days on specialised hardware.

One way to reduce the marginal costs is to have central data repositories with select CDMs that underpin pharmacovigilance applications, so the added costs are limited to running the analyses. We partook in a proof-of-concept study evaluating the OMOP CDM with use cases in patient-level prediction and pharmacovigilance. Although we, regrettably, never got to publish any results,[34] the

[34] Apart from a poster at the European OHDSI Symposium 2019 [11]

experience yielded two learning points: setting up a sound extract-transform-load (ETL) process is a demanding task, but once it is done and validated the opportunities do make up for it.[35]

[35] See also e.g. the European Medicines Agency's recently closed call for tenders with data in the OMOP CDM

The future

Below I point to 4 avenues that seem particularly apt for future work at the nexus between secondary observational data, machine learning, and pharmacovigilance.

Deep Danish data in common data models

We have excellent health and administrative data in Denmark, but due to the relatively little population, we may not have enough observed events of potential ADRs to have sufficient power for safety signal detection or aetiological studies. This renders transnational collaboration necessary and our data infrastructure should facilitate this.

Having all Danish observational data available out-of-the-box in appropriate CDMs would facilitate faster pharmacovigilance enquiries with more power through e.g. OHDSI network studies, help overcome the issue of opague data processing and entailing duplicative efforts, and enable leveraging the full Danish population with deeper and better data than can currently be used for pharmacovigilance. Indeed, a recent the European Medicines Agency (EMA) guideline on register-based studies lists OMOP as a recommended terminology[36] [184, appendix 3]. This notion is not new in a Danish setting: Aarhus University Hospital participates in the EU-ADR alliance [185].

[36] Although OMOP is not strictly a terminology but a CDM that uses standard vocabularies to achieve harmonisation

We cannot compete with North American databases with hundreds of millions of patients [186], but we can have deeper data for better patient characterisation. As discussed previously, this could be e.g. nurses' notes that provide other aspects and closer-to-real-time data on patient status, or human microbiome data due to its (potential) interference with drug uptake and response [187, 188].

Multilingual safety signal detection

As we illustrate with study III, there is scope for using language-agnostic text mining in clinical notes for safety signal detection. In particular, our approach might lend itself well to transnational text mining using machine translation.

This could be a viable complement to current efforts usign natural language processing (NLP) for pharmacovigilance, but its real potential remains to be explored and substantiated. Although less than for other types of studies using structured data, even for this kind of endeavours, using CDMs would be convenient as it would simplify the development of the analytic code to set up the system locally.

Machine learning-based propensity scores

Genuinely useful medical applications of machine learning have hitherto been mainly in data-rich domains with the kind of data these methods were built for: images (radiology [189], ophtalmology [190], pathology [191]) and intricate, high-dimensional time-series data (intensive care medicine [106, 126]). But medicine overall has turned out to not be as low-hanging a fruit as anticipated, perhaps with Watson for Oncology as the most spectatular example [192].

Algorithmic bias is one very important thing to keep in mind in this regard. This comes back to the challenges presented in figure 32: a key source of algorithmic bias is actually the well-known selection bias, that is, if the data are not representative of the target population, the results will be biased and sometimes in unpredictable ways, and bigger data set will not fix this automatically.[37]

[37] "Running a poorly designed algorithm on a faster computer doesn't make the algorithm better; it just means you get the wrong answer more quickly. (And with more data there are more opportunities for wrong answers!)" Russell [78, p. 37]

Causal inference based on propensity-score modelling is a promising avenue for using machine learning models in pharmacovigilance. In the OHDSI toolbox, propensity scores are based on lasso logistic regression, but we might be better off with more complex machine learning models [193], especially for deep data as we might recoup more information this way than is (hopefully) captured by hundreds or thousands of features in a penalised (but linear nonetheless) logistic regression.

Some results suggest that using propensity scores obtained from more complex machine learning models yield superior results [194], but it will be important to submit such models to rigorous evaluation to ascertain adequate predictive performance.

Safety signal detection + causal discovery

Causal discovery (also known as structure learning) allows for inferring causal mechanisms from observational data [153, 195] and has obvious applications in medical research, as an alternative to causal inference methods based on e.g. propensity scores.

Important statistical characterisics are still largely unexplored, somewhat precluding wider uptake for causal inference. Despite this, using causal discovery in observational data could be a good way to complement safety signal detection in ICSRs in safety signal triangulation [196], as it is more data-driven than conventional causal inference building on prespecified hypotheses.

III

Prevalence and adverse outcomes of drug-drug interactions

An earlier version of this manuscript was included in the PhD thesis of Cristina Leal Rodríguez [7].

Full title

Drug interactions in hospital prescriptions in Denmark: Prevalence and associations with adverse outcomes

Chapter contents

Drug interactions in hospital prescriptions in Denmark: Prevalence and associations with adverse outcomes

Cristina Leal Rodríguez MSc[1,§], Benjamin Skov Kaas-Hansen MD MSc[1,2,§], Robert Eriksson MD PhD[1,3], Jorge Hernansanz Biel BSc[1], Kirstine G. Belling PhD[1], Stig Ejdrup Andersen MD PhD[2], Søren Brunak PhD[1,*]

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark

[2]Clinical Pharmacology Unit, Zealand University Hospital, DK-4000 Roskilde, Denmark

[3]Department of Pulmonary and Infectious Diseases, Nordsjællands Hospital, DK-3400 Hillerød, Denmark

§ Joint first authors.

* To whom correspondence should be addressed: soren.brunak@cpr.ku.dk, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark

Abstract

Importance: While the beneficial effects of medications are numerous, drug-drug interactions may lead to adverse drug reactions that are preventable causes of morbidity and mortality.

Objective: To quantify the prevalence of potential drug-drug interactions in drug prescriptions at Danish hospitals, estimate the risk of adverse outcomes associated with discouraged drug combinations, and highlight the patient types (defined by the primary diagnosis of the admission) that appear to be more affected.

Design: Cross-sectional (descriptive part) and cohort study (adverse outcomes part).

Setting: Hospital electronic health records from two Danish regions (approx. 2.5 million people) from January 2008 through June 2016.

Participants: Inpatients receiving two or more medications during their admission.

Exposure: Concomitant prescriptions of potentially interacting drugs as per the Danish Drug Interaction Database.

Main outcome and measure: Descriptive part: prevalence of potential drug-drug interactions in general and discouraged drug pairs in particular during admissions. Adverse outcomes part: post-discharge all-cause mortality rate, readmission rate and length-of-stay.

Results: Among 2,886,227 hospital admissions (945,475 patients; median age 62 years [IQR: 41-74]; 54% female; median number of drugs 7 [IQR: 4-11]), patients in 1,836,170 admissions were exposed to at least one potential drug-drug interaction (659,525 patients; median age 65 years [IQR: 49-77]; 54% female; median number of drugs 9 [IQR: 6-13]), and in 27,605 admissions to a discouraged drug pair (18,192 patients; median age 68 years [IQR: 58-77]; female 46%; median number of drugs 16 [IQR: 11-22]). Meropenem-valproic acid (HR: 1.5, 95% CI: 1.1–1.9), domperidone-fluconazole (HR: 2.5, 95% CI: 2.1–3.1), imipramine-terbinafine (HR: 3.8, 95% CI: 1.2–12), agomelatine-ciprofloxacin (HR: 2.6, 95% CI: 1.3–5.5), clarithromycin-quetiapine (HR: 1.7, 95% CI: 1.1–2.7), and piroxicam-warfarin (HR: 3.4, 95% CI: 1–11.4) were associated with elevated mortality. Confidence interval bounds of pairs associated with readmission were close to 1; length-of-stay results were inconclusive.

Conclusions and Relevance: Well-described potential drug-drug interactions are still missed and alerts at point of prescription may reduce the risk of harming patients; prescribing clinicians should be alert when using strong inhibitor/inducer drugs (i.e. clarithromycin, valproic acid, terbinafine) and prevalent anticoagulants (i.e. warfarin and NSAIDs) due to their great potential for dangerous interactions. The most prominent CYP isoenzyme involved in mortality and readmission rates was 3A4.

Introduction

Two drugs are said to interact when the action of one does or may affect the activity, metabolism or toxicity of the other[1]. Drug-drug interactions (DDIs) constitute a particularly important cause of adverse drug reactions (ADRs) as clinical evidence and (when known) their pharmacological mechanisms make them somewhat predictable. Many hospitalised patients take several drugs and polypharmacy[2] is estimated to affect 40–65% of hospitalised patients[3,4].

Although the risk of DDIs is proportional to the number of drugs taken[5], the clinical consequences vary widely, and ADRs rarely occur[6]. Even if uncommon, serious adverse outcomes do cause harm, constitute economic losses and are to some extent preventable. At particularly elevated risk of ADRs are the elderly (often multimorbid and with reduced physiological capacity)[2] and patients with diseases in organ systems involved in drug metabolism, particularly kidneys and liver[7]. The consequences of DDIs affect both the individual patient and society as a whole: 10–20% of hospital admissions may be attributable to drug-related problems and toxic effects of medication of i.a. DDIs[8,9], and studies have linked DDIs to prolonged hospitalisation and increased healthcare costs[10-13].

The electronic medication management systems deployed at public hospitals in Denmark do not systematically flag problematic drug combinations. Even with such systems in place, alert fatigue is a real issue that requires tailoring to optimise their genuine utility[14]. To this end, appropriate evidence about the extent and nature of the problem is needed.

No studies to date have examined the prevalence of potential drug-drug interactions (pDDIs) in hospitals for different patient types and assessed the clinical impact of pDDIs. This study sought to fill this gap and elicit learning points for clinicians to mitigate this issue.

We used electronic health records (EHRs) to (a) elicit the prevalence of discouraged drug pairs and their expected clinical significance and documentation level, (b) identify which patient types are most affected by discouraged pairs, and (c) gauge the association between discouraged pairs and three adverse outcomes: post-discharge mortality, readmission, and length-of-stay (LOS).

Materials and methods

Patients and data

We obtained inpatient data for admissions to twelve public hospitals in the Capital Region and Region Zealand, Denmark, from January 2008 through June 2016. The two regions comprise approximately 2.5 million people, about half of the Danish population[15]. Admissions of individuals using at least two drugs concomitantly were included. We defined concomitant use as temporally overlapping time prescriptions and identified all two-way drug combinations.

Information on admission timing, diagnoses, and medical histories was obtained from the Danish National Patient Register (DNPR)[16,17], recording data for department-specific visits. DNPR encodes diagnoses with a Danish version of the International Classification of Disease, 10th revision (ICD-10). An admission's primary diagnoses are recorded retrospectively at discharge. Successive in-hospital visits were combined into admissions if they were at most one day apart.

We marshalled information on dispensed in-hospital drug prescriptions from OPUS-medication (OpusMed) and Electronic Patient Medication (EPM). The latter has been validated[18] and the former was used in the same manner; both use the WHO Anatomical Therapeutic Chemical (ATC) classification system[19].

As our pDDI reference we used the Danish Drug Interaction Database (DID), covering predominantly pharmacokinetic interactions based mainly on published results and maintained by specialists in clinical pharmacology under the auspices of the Danish Medicines Agency[20].

pDDI prevalence

This descriptive part was cross-sectional. pDDIs were categorised by management recommendation (five levels), clinical significance (five levels), and documentation level (six levels); we only considered the 14,237 (from a total of 18,691) pDDIs with information on all three axes (**Table 1**). The quality of the documentation level is based on the evidence about the significance of the kinetic or dynamic properties.

Discouraged drug pairs were defined as prevalent when they occurred in more than 10% of admissions of at least one specific patient type, defined as the ICD-10 chapter of the

admission's primary diagnosis. We used standardised difference in proportions to compare imbalances between binary variables, taking an absolute difference above 10% to indicate substantial imbalance[21].

Adverse outcomes of exposure to discouraged combinations

In this analytic part of the study, we screened the effect of all discouraged pairs on post-discharge all-cause mortality rate (henceforth, post-discharge mortality), readmission rate and LOS. Only patients' first admissions were used. We excluded patients whose exposure started outside the hospital for better-defined exposure start. The effects on post-discharge mortality and readmission were estimated with stratified Cox regression models assuming noninformative censoring[22] and the effects on LOS with stratified Poisson regression models[23], with exposure to the discouraged drug pair as the sole explanatory variable. We created strata by greedy 1:5 matching on preference score, an extension of the propensity score accounting for target exposure prevalence[24]. The preference score is the probability that a patient be exposed whether this happened or not. Thus, if two patients have (almost) the same preference score but one was exposed and the other not, the exposure is a likely explanation for their difference in outcome[25-27].

We used Cyclops[28] to compute high-dimensional propensity scores[26] with sparse lasso logistic regression models using up to 843 features derived from eight covariates: age at admission (continuous), sex (binary), patient type (one-hot-encoded), diagnoses during admission (ICD-10 level 3, one-hot-encoded), medication burden (continuous), whether the admission was acute or elective (binary), and weighted Elixhauser comorbidity score (Agency for Healthcare Research Quality[29] version, continuous). Seeking empirical equipoise, outcome models were fit to patients with preference scores between 0.3 and 0.7[24]. The significance level was set to 5%; power analyses were foregone. Estimates with 95% confidence intervals (CI) wider than 100 on the linear scale were omitted.

Software

We used the R statistical programming language and Python for data processing, analysis, and visualisation. The analysis workflow was built as a Snakemake pipeline[30] (**eFigure 1)**. The full analytic code is available upon request.

Ethics

Data were stored and analysed on a secure cloud in Denmark. Registry data access was approved by the Danish Health Data Authority (FSEID-00003092, FSEID-00004491, FSEID-00003724) and the Danish Patient Safety Authority, which at the time was the competent body for approvals regarding research in EHRs, approved journal access and the purpose for the study (3-3013-1731-1).This article observes relevant items in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[31].

Results

Among the 4,411,576 admissions of 1,481,584 patients identified, we included 2,886,227 admissions (65%) of 945,475 patients (64%) to whom two or more drugs were administered (**eFigure 2**). **Table 2** shows overall and stratified summary statistics for pertinent variables. The 538,620 (57%) women in the cohort contributed 1,551,131 admissions (54%) and 13,122,610 (54%) dispensed prescriptions. Of these, 27,605 admissions (1%) featured discouraged drug pairs and 12,655 (46%) were administrated to women. pDDIs and discouraged drug pairs were observed more frequently in older patients. Further, the median number of prescribed drugs in admissions with discouraged drug pairs (16, IQR: 11-22) was larger than any-pDDI (9, IQR: 6-13) and no-pDDI (4, IQR: 2-6) admissions. Patients exposed to discouraged drug pairs were more ill and had longer admissions and higher in-hospital mortality.

Of 344,489 unique drug pairs administered in-hospital, 5,646 (2%) were pDDIs; 1,836,170 admissions (64%) of 659,525 patients (70%) featured at least one of these 5,646 pDDIs. In 27,605 admissions (1%) of 18,192 patients (2%) at least one of the 146 (3%) discouraged drug pairs was used, most with expected major (71%) and moderate (21%) clinical significance (**Table 3** and **eTable 1**). The most prescribed drugs involved in discouraged drug pairs were, in descending order of number of users, pantoprazole (nine admissions [0.0%] of five patients [0.0%] exposed to discouraged drug pairs of 570,440 admissions of 224,002 pantoprazole users), ibuprofen (9,982 admissions [1.8%] of 7,368 patients [2.0%] of 569,223 admissions of 365,302 users), simvastatin (5,048 admissions [1.1%] of 3,887 patients [3.6%] of 442,545 admissions of 148,579 users), metoprolol (1,191 admissions [0.3%] of 399 patients [0.3%] exposed of 379,785 admissions of 127,237 users), and

diclofenac (1,917 admissions [1.1%] of 1,326 patients [1.1%] exposed of 177,928 admissions of 120,256 users) affecting up to 3% of the hospitalized patients receiving the drugs (**eTable 2**). In contrast, more uncommon drugs (used by less than 1% of hospitalized patients), e.g. erythromycin (1,573 admissions [27.8%] of 1,461 patients [29.2%] exposed out of 5,665 admissions of 5,001 users), rifabutin (25 admissions [24.8%] of 10 patients [21.7%] exposed out of 101 admissions of 46 users), ketoconazole (644 admissions [20.4%] of 320 patients [21.2%] exposed out of 3,158 admissions of 1,513 users), warfarin (12,570 admissions [10.3%] of 8,791 patients [20.9%] exposed out of 121,653 admissions of 42,101 users), and domperidone (2,872 admissions [12.4%] of 2,028 patients [19.2%] exposed out of 23,213 admissions of 10,571 users) were more often given as part of discouraged pairs (**eTable 2**, **eFigure 3**).

Overall, patients admitted with cardiovascular diseases (ICD-10 chapter IX); endocrine, nutritional and metabolic diseases (chapter IV); and respiratory diseases (chapter X) were more frequently exposed to discouraged drug pairs unlike obstetrical patients (chapter XV) and patients admitted for other reasons (chapter XXI) (**Figure 1A**, **eFigure 4**). Discouraged pairs varied among the remaining patient types, but within the ±0.1 threshold indicative of negligible imbalance (**Figure 1A**). In contrast, most drugs were more frequently prescribed in admissions with discouraged pairs with many above the 0.1 threshold except misoprostol and oxytocin (**Figure 1B**).

In the 65 discouraged drug pairs (45%) prescribed to five patients or more (**eTable 3**), seven were prevalently (>10% of admissions) prescribed during hospital admissions (**Figure 2**). The most prominent pair was warfarin-ibuprofen, prevalent in all patient types except three (chapters X, XVI and XX). The second-most prominent was simvastatin-clarithromycin, prevalent in six patient types (I, III, IV and X-XII); the third-most was domperidone-fluconazole, prevalent in five patient types (II-IV, VXIII and XXI). The other four were warfarin-diclofenac (XIV, XV, XVII), fluoxetine-venlafaxine (V), meropenem-valproic acid (VI) and erythromycin-fluconazole (XI). **eFigures 4-6** show the prevalence of each drug and each diagnosis in patients exposed vs non-exposed to discouraged drug pairs.

Figure 3 shows the estimated effects of exposure on mortality rate, readmission rate and LOS; **eTable 5** contains the numerical estimates. Six discouraged drug pairs were significantly associated with increased mortality rate, of which particularly the 95% CIs of

meropenem-valproic acid, domperidone-fluconazole, imipramine-terbinafine and agomelatine-ciprofloxacin are relatively far from 1. Ertapenem-fluconazole, amitriptyline-terbinafine as well as clarithromycin with ticagrelor, tacrolimus and everolimus, respectively, were associated with substantially elevated readmission rates albeit with CI bounds near 1. Many discouraged pairs were associated with longer or shorter hospital stays with most effect sizes within approximately ±1 day.

Discussion

We found that 1,836,170 admissions (64%) of 659,525 patients (70%) featured at least one pDDI and that during 27,605 admissions (1%) of 18,192 patients (2%) at least one discouraged drug pair was administered. Seven discouraged pairs were prevalent, most notably warfarin-ibuprofen (18 patient types), simvastatin-clarithromycin (six patient types) and domperidone-fluconazole (five patient types). Of the prevalent discouraged pairs, domperidone-fluconazole and meropenem-valproic acid (one patient type) were significantly associated with elevated mortality. The prevalent pair warfarin-ibuprofen was just statistically significantly associatiated with elevated readmission rates and three of five discouraged pairs associated with elevated readmission rates involved clarithromycin. LOS results were inconclusive.

The increasing availability of longitudinal patient data and growing access to databases with DDI information facilitate comparative, data-driven approaches to identify, anticipate and explain DDIs[32]. Indeed, in this study we used comprehensive phenotypic in-hospital data to detail the landscape of in-hospital pDDIs with particular focus on discouraged drug pairs to elicit their effects on potentially preventable adverse outcomes.

Prevalence patterns and effects of pDDIs are elusive because many potentially interacting drug combinations offer genuine clinical utility if used consciously by alert physicians. Teasing apart these dynamics is difficult on a large scale. Indeed, studies of pDDI prevalence in hospitalised patients often use relatively small samples from sub-populations such as critically ill or oncological patients[33-35]. Our approach was different seeking to conduct a large-scale screening of hospitalised patients, focusing on outright discouraged drug pairs because their clinical benefits unlikely outweigh their potential harm.

A recent systematic review of clinically manifested DDIs[36] found prevalence estimates up to 26% in not-critically-ill hospitalised patients[37]. Further, the number of drugs used concomitantly has been shown to be a significant risk factor for interactions at hospitals and in primary care[38-41]. We also observed widespread polypharmacy among patients exposed to pDDIs especially when exposed to discouraged drug pairs. However, unlike for diagnoses, no drugs involved in pDDIs emerged as neither particularly frequent nor infrequent except misoprostol and oxytocin. Thus, perceiving the effect of polypharmacy solely in terms of the association between number of concomitant drugs and pDDIs is arguably of limited use as it tells us little about the nature of the association. Instead, other phenotypic factors such as comorbidities may be of greater utility to the prescribing physician at point of care.

A Danish study of 167,232 patients from 1998 on the island of Funen found that 4.4% of all inhabitants of age above 70 were prescribed drug combinations with a high risk of severe interactions[42]. A recent Brazilian study with approximately 340,000 patients from primary- and secondary-care hospitals arrived at a similar figure[43]. These estimates are substantially lower than our 14% patients prescribed pDDIs with expected major clinical significance (**Table 3**), likely because our data are newer than those in the former and include also tertiary hospitals unlike both those studies.

Another study from Denmark published in 2005 found that pDDIs are prevalent but mostly clinically insignificant[44]. Our results agree with this notion: six discouraged combinations featured substantial and statistically significant associations with elevated mortality, of which only two were prevalent in particular patient types (meropenem-valproic acid, domperidone-fluconazole). This was the case for only warfarin-ibuprofen with respect to readmission rates.

Rarely used drugs are more often involved in potentially dangerous DDIs perhaps due to prescribers' lack of specific knowledge on these drugs; consider three examples. First, using meropenem (or ertapenem) with valproic acid elevates the risk of seizures (unknown mechanism) and meropenem consumption is increasingly prescribed at emergency departments, often by junior doctors. Second, the cardiac risks of domperidone and erythromycin (prolonged QT and Torsades-de-Pointes) are aggravated by concurrent use of fluconazole (or any conazole) because the latter impedes their metabolism by inhibiting CYP3A4[45]. Third, concurrent use of agomelatine and ciprofloxacin increases the exposure of

the former because the later inhibits CYP1A2. Causes of death in deceased exposed to these drug pairs did not suggest unexpected patterns (**eFigure 7**).

Some active ingredients involved in discouraged drug pairs have several ATC codes (terbinafine, diclofenac and tacrolimus) and the somewhat agreeing effect estimates on mortality and readmission rates add confidence to these findings. Interestingly, the effects on LOS were not consistent across ATC codes for the same active ingredient prompting cautious interpretation. Indeed, LOS is elusive: for example, a short admission can end with discharge to home or death. To arrive at meaningful conclusions on lengths-of-stays, one would need to use for example drug administrations allowing for time-to-event analyses, something not possible with these data.

Strengths and limitations

This study features a range of strengths. First, this is the largest study assessing the prevalence of discouraged drug pairs and their effects on adverse outcomes among hospitalised patients. Second, we used unfiltered data from a heterogeneous population of almost one million hospitalised patients over an eight-year period. Third, detailed and reliable register data allow for detailed phenotyping, both with respect to diagnoses and medication use. Fourth, such deep phenotyping underpins the use of high-dimensional preference scores to obtain approximate empirical equipoise when studying the associations between exposure and adverse outcomes. Fifth, the risk of selection bias and loss to follow-up was minimal.

Nonetheless, there are potential weaknesses. First, hospital data may be subject to recall and information bias. This is likely not an issue for this study because we rely on near-objective data (e.g. validated source of medication data) used also for administrative and billing purposes. Bias by indication could be a problem but the use of high-dimensional propensity scores should, at least in part, counter this. Second, we only considered two-way pDDIs. Large-scale screening for N-drug interactions is difficult due to combinatorial explosion in the number of possibilities and difficulties in defining a proper reference to which the results should be compared. Instead, targeted investigations would be meaningful, e.g. on triple whammy and its effect on kidney function. Third, different pDDI databases likely feature discrepancies regarding management recommendations and clinical significance, and the DID covers primarily pharmacokinetic interactions. Further, DID allows different levels of evidence: for older drugs only pDDIs supported by published evidence are considered,

whereas for newer drugs also pDDIs from summaries of product characteristics not published elsewhere are included. This database, nonetheless, is well-known among Danish physicians and used in daily practice to guide medicinal treatment and, as such, makes for a natural gold standard against which to compare real-life prescriptions in Denmark. Fourth, the pDDIs involving antibiotics and systemic antifungals and associated with elevated mortality are used to treat serious infections. Thus, exposure to these combinations could be proxies for serious clinical conditions, themselves associated with high mortality. If so, physicians could have deemed it worthwhile to use a discouraged drug pair due to bleak prognoses. Fifth, despite a large dataset we had relatively few patients exposed to several discouraged drug pairs, making it difficult to rule out effects of these exposures on mortality and readmission rates even though we did not find any.

Conclusion

Discouraged drug pairs are common in hospitalised patients at large and so are potentially problematic drug pairs, notably, combinations of warfarin and NSAIDs and with antiinfectives (especially, azoles, carbapenems and macrolides). The meropenem-valproic acid and domperidone-fluconazole combination, both prevalent in at least one patient type, were significantly associated with elevated post-discharge mortality rate. This study elicited unfortunate prescription patterns with potentially detrimental effects in hospitalised patients and the CYP3A4 isoenzyme was involved in more than half the discouraged pairs associated with elevated mortality or readmission rates.

Competing interests

S.B. reports ownerships in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S and managing board memberships in Intomics A/S outside the submitted work. All other authors report no competing interests.

Acknowledgements

R.E. proposed the idea. C.L.R. and B.S.K.H. designed and conducted the study. S.B. and S.E.A. co-created the project that made available and curated the EHR data. S.B. and S.E.A. obtained funding. C.L.R., B.S.K.H. and J.H.B. performed pre-processing of the data. (B.S.K.H. extracted eGFR measurements, and calculated comorbidity scores; J.H.B. extracted the information from Danish interaction database under the supervision of C.L.R. C.L.R pre-processed the clinical and medication data, calculated the treatment exposures and

treatment overlaps). C.L.R. and B.S.K.H. performed the computational and statistical analysis. B.S.K.H. did the adverse outcome modelling. C.L.R. and B.S.K.H. interpreted data and provided critical intellectual content, under guidance by S.E.A. C.L.R. performed the literature search. C.L.R. and B.S.K.H. wrote the initial draft. All authors have contributed to and approved the final manuscript.

This work was supported by the Novo Nordisk Foundation (grants NNF14CC0001 and NNF17OC0027594) and the Danish Innovation Fund (grant 5153-00002B). The funding bodies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. C.L.R and B.S.K.H. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Tables

Table 1. Classification of potential drug-drug interactions based on management recommendation, clinical significance, and documentation level published by the Danish Medicines Agency Drug Interaction Database.

Management recommendation	
1	The drug combination should always be avoided (discouraged in text).
2	The drug combination can be used with dose adjustment.
3	The drug combination can be used with staggered time of ingestion.
4	The drug combination can be used under certain precautions, i.e. changing the routes of administration. Alternative agents should be considered.
5	The drug combination can be used. No action needed as the risk of adverse events appears to be small.
Clinical significance	
Major	Clinically pronounced/physiological effect with either significant altered therapeutic response or frequent occurrence of serious adverse reactions.
Moderate	Clinically moderate/physiological effect with either slightly altered therapeutic response, or rare occurrence of more serious side effects. Serum concentration changes, which in other experiments have been closely associated with the above-mentioned phenomena.
Minor	Unchanged or not significantly altered biological response with fewer and easier side effects - or serum concentration changes, which in other studies have not shown significant changes in the biological response.
Possible	Pharmacokinetic changes which are not accompanied by known adverse reactions or changes in the biological response.
None	Neither kinetic or physiological/clinical changes.
Undetermined	Kinetic or physiological/clinical changes that cannot be estimated based on the available documentation.
Documentation level	
Well-documented	At least 2 (from different centres) human controlled trials and/or (before and after) trials in relevant individuals with single or multiple steady state trials in the form of either significant kinetic or dynamic changes.
Documented	A human controlled study and/or (before and after) study with steady state single or multiple dose trials in the form of either significant kinetic or dynamic changes.
Limited documented	Either more than 2 case reports with relevant during and after kinetics or dynamics or human *in vitro* studies with relevant cytochrome P450 (CYP) fractions and concentrations.
Poorly documented	1-2 case reports. Non-conclusive *in vitro* studies.

Table 2. Overall and stratified summary statistics of included admissions. Values are N (%) and median (interquartile range). pDDI: potential drug-drug interaction. AHRQ: Agency for Healthcare Research Quality.

	Overall	No pDDIs	pDDIs	Discouraged drug pairs
Admissions	2,886,227	1,050,057 (36%)	1,836,170 (64%)	27,605 (1%)
Women	1,551,131 (54%)	565,697 (54%)	985,434 (54%)	12,655 (46%)
Patients	945,475	553,612	659,525	18,192
No. prescriptions	9 (5-15)	5 (3-8)	12 (7-19)	22 (14-36)
in women	8 (4-15)	4 (3-7)	12 (7-19)	22 (14-35)
No. unique prescribed drugs	7 (4-11)	4 (2-6)	9 (6-13)	16 (11-22)
Unique prescribed drugs				
2-4 drugs	888,934 (31%)	629,786 (60%)	259,148 (14%)	520 (2%)
5-9 drugs	1,042,023 (36%)	347,943 (33%)	694,080 (38%)	4,408 (16%)
≥ 10 drugs	955,270 (33%)	72,328 (7%)	882,942 (48%)	22,677 (82%)
Age in years	62 (41- 74)	51 (30-69)	65 (49-77)	68 (58-77)
Age group				
< 18 years	203,125 (7%)	148,043 (14%)	55,082 (3%)	492 (2%)
18-44 years	619,540 (22%)	293,005 (28%)	326,535 (18%)	2,359 (9%)
45-64 years	773,558 (27%)	269,693 (26%)	503,865 (27%)	7,790 (28%)
65-74 years	577,389 (20%)	162,494 (16%)	414,895 (23%)	8,166 (30%)
75-84 years	461,247 (16%)	114,094 (11%)	347,153 (19%)	6,588 (24%)
≥ 85 years	251,368 (9%)	62,728 (6%)	188,640 (10%)	2,210 (8%)
pDDIs per patient	1 (0-3)	0 (0-0)	2 (1-5)	9 (5-15)
Length of stay in days	3 (1-6)	2 (1-4)	3 (2-7)	7 (3-15)
Acute admission	2,107,774 (73%)	765,816 (73%)	1,341,958 (73%)	19,746 (72%)
In-hospital mortality	62,830 (2%)	14,397 (1%)	48,433 (3%)	1,252 (5%)
Low eGFR (<30ml/min/1.73m^2)	109,907 (4%)	15,198 (1%)	94,709 (5%)	2,660 (10%)
Elixhauser index (AHQR)				
<0	646,561 (22%)	230,311 (22%)	416,250 (23%)	4,893 (18%)
0	854,868 (30%)	399,526 (38%)	455,342 (25%)	3,838 (14%)
1-4	297,174 (10%)	104,072 (10%)	193,102 (11%)	3,157 (11%)
≥5	1,087,624 (40%)	316,148 (30%)	771,476 (42%)	15,717 (57%)
Most common drug classes (ATC level 3)				
Other analgesics and antipyretics (N02B)	1,334,677 (63%)	501,208 (48%)	1,334,677 (73%)	22,024 (80%)

Antithrombotic agents (B01A)	1,038,880 (43%)	211,944 (20%)	1,038,880 (57%)	21,890 (79%)
Opioids (N02A)	917,092 (43%)	318,180 (30%)	917,092 (50%)	18,098 (66%)
Anti-inflammatory and antirheumatic products, non-steroids (M01A)	663,518 (29%)	178,152 (17%)	663,518 (36%)	14,900 (54%)
Drugs for peptic ulcer and gastro-oesophageal reflux disease (GORD) (A02B)	642,650 (27%)	141,344 (13%)	642,650 (35%)	15,429 (56%)
Beta-lactam antibacterials, penicillins (J01C)	686,899 (24%)	206,305 (20%)	480,594 (26%)	10,796 (39%)
Loop (high-ceiling) diuretics (C03C)	518,342 (18%)	52,487 (5%)	465,855 (25%)	131,98 (48%)
Most common primary diagnosis				
Abdominal and pelvic pain (R10)	63,574 (2%)	26,387 (3%)	37,187 (2%)	448 (2%)
Pneumonia, organism unspecified (J18)	60,237 (2%)	21,369 (2%)	38,868 (2%)	1,057 (4%)
Atrial fibrillation and flutter (I48)	55,405 (2%)	14,422 (1%)	40,983 (2%)	667 (2%)
Mental and behavioural disorders due to use of alcohol (F10)	51,347 (2%)	29,206 (3%)	22,141 (1%)	135 (0%)
Other chronic obstructive pulmonary disease (J44)	45,919 (2%)	16,039 (2%)	29,880 (2%)	474 (2%)
Nonrheumatic aortic valve disorders (I35)	10,244 (0%)	1,784 (0%)	8,646 (0%)	1,439 (5%)
Acute myocardial infarction (I21)	30,250 (1%)	983 (0%)	29,267 (2%)	244 (1%)
Angina pectoris (I20)	31,664 (1%)	5,366 (1%)	26,298 (1%)	190 (1%)
Bacterial pneumonia, NOC (J15)	24,045 (1%)	8,420 (1%)	15,625 (1%)	528 (2%)
Other sepsis (A41)	28,401 (1%)	7,348 (1%)	21,053 (1%)	482 (2%)

Table 3. Unique drug combinations (upper cells) and prevalence (lower cells) of pDDIs by management recommendation and clinical significance. Values are N (%).

Recommendation level	Clinical significance						
	Major	Moderate	Minor	Possible	None	Un-determined	Total
1: Discouraged	104 (71)	31 (21)	-	8 (5)	-	3 (2)	146 (3)
	16,339 (90)	1,293 (7)	-	1,206 (7)	-	24 (0)	18,192 (3)
2: Dose adjustment	164 (16)	457 (45)	48 (5)	279 (28)	1 (0)	56 (6)	1,005 (18)
	40,718 (27)	91,264 (61)	12,606 (8)	58,622 (39)	25 (0)	4,953 (3)	148,455 (23)
3: Staggered ingestion	53 (23)	100 (44)	9 (4)	47 (21)	-	17 (8)	226 (4)
	14,339 (16)	38,544 (43)	244 (0)	12,264 (14)	-	43,776 (48)	90,662 (14)
4: Precautions	300 (17)	602 (35)	86 (5)	601 (35)	30 (2)	123 (7)	1,742 (31)
	45,221 (8)	459,717 (86)	82,611 (16)	249,539 (47)	12,214 (2)	106,406 (20)	532,066 (81)
5: No action needed	6 (0)	82 (3)	311 (12)	206 (8)	1,648 (65)	274 (11)	2,527 (45)
	165 (0)	97,285 (21)	189,797 (40)	116,185 (25)	399,935 (85)	191,767 (41)	470,956 (71)
Total	627 (11)	1,272 (23)	454 (8)	1,141 (20)	1,679 (30)	473 (8)	5,646
	92,167 (14)	517,599 (78)	225,512 (34)	1,679 (0)	400,935 (61)	264,711 (40)	659,525

Figures

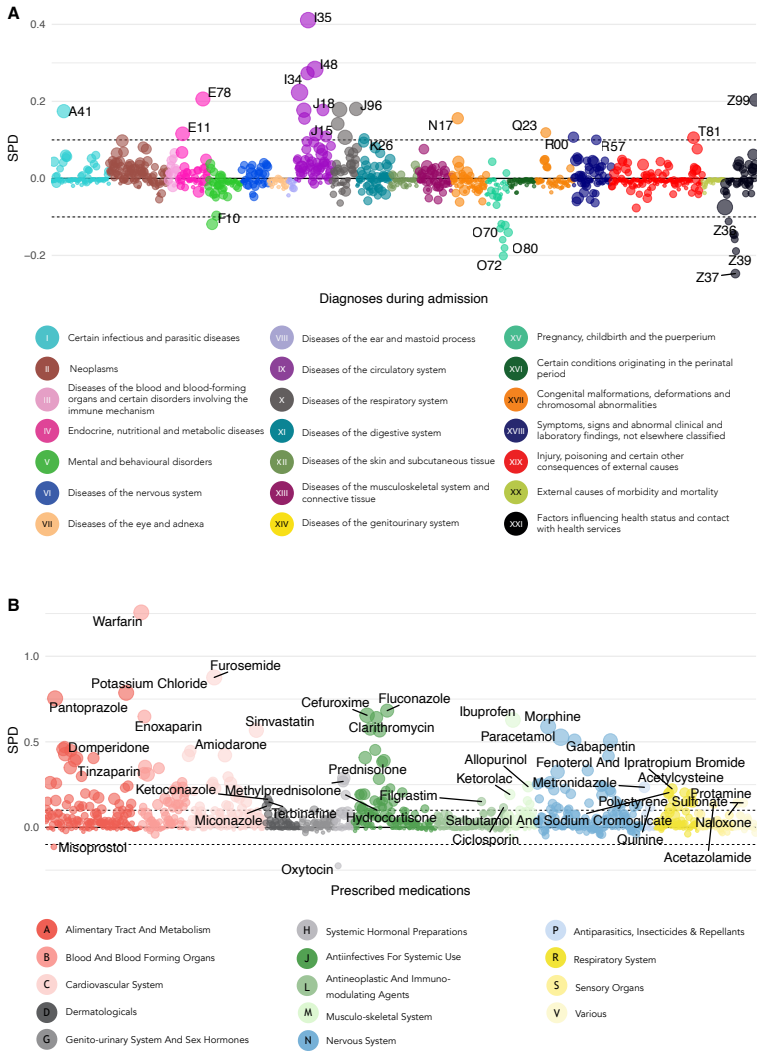


Figure 1. Standardised differences in proportions (i.e. discouraged drug pairs initiated versus not) of diagnoses (panel A) and prescribed drugs during admissions (panel B), respectively. The colour represents ICD-10 chapter and anatomical ATC level, respectively, and the size is the prevalence in patients exposed to discouraged drug pairs. The top three diagnoses and drugs are labelled, and an interactive version of the figure is provided as online supplementary material.

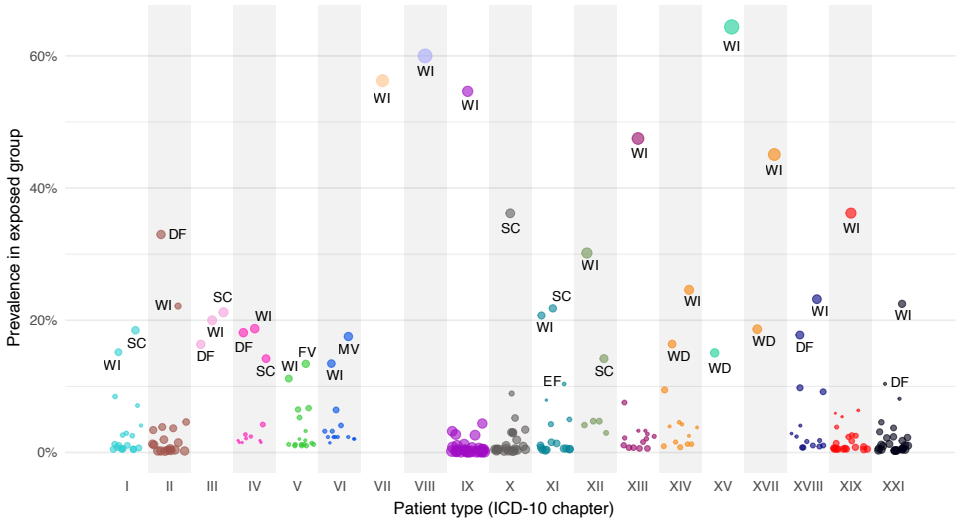


Figure 2. Prevalence of discouraged drug pairs by patient type. Each point represents one discouraged drug pair, and size the absolute value of the standardised difference in proportions using as reference admissions during which treatment with any discouraged pair was initiated. DF (N = 5): Domperidone (A03FA03) + Fluconazole (J02AC01); WD (N = 3): Warfarin (B01AA03) + Diclofenac (M01AB05, systemic); WI (N = 18): Warfarin (B01AA03) + Ibuprofen (M01AE01); SC (N = 6): Simvastatin (C10AA01) + Clarithromycin (J01FA09); MV (N = 1): Meropenem (J01DH02) + Valproic acid (N03AG01); EF (N = 1): Erythromycin (J01FA01) + Fluconazole (J02AC01); FV (N = 1): Fluoxetine (N06AB03) + Venlafaxine (N06AX16).

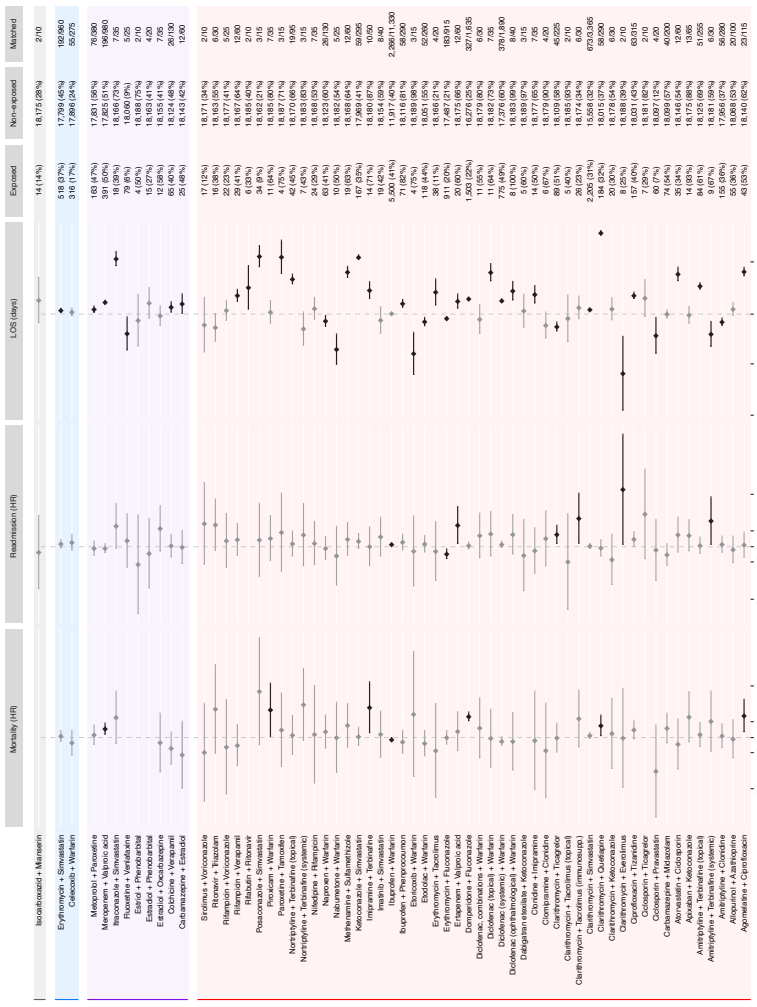


Figure 3. Estimate effect sizes of exposure to discouraged drug pairs and post-discharge mortality rate (hazard ratio, HR), readmission rate (HR) and length-of-stay (change in days). Diamonds show point estimates of the effect sizes, horisontal lines the 95% confidence intervals. The exposed and non-exposed columns show count (empirical equipoise) and the matched column shows the number of exposed/non-exposed used to estimate the effects of that pair.

1 References

1. Baxter K, Preston CL. *Stockley's drug interactions.* Vol 495: Pharmaceutical Press London; 2010.
2. Guthrie B, Makubate B, Hernandez-Santiago V, Dreischulte T. The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010. *BMC Medicine.* 2015;13(1):74-74.
3. Doucet J, Chassagne P, Trivalle C, et al. Drug-drug interactions related to hospital admissions in older adults: a prospective study of 1000 patients. *Journal of the American Geriatrics Society.* 1996;44(8):944-948.
4. Egger SS, Drewe J, Schlienger RG. Potential drug–drug interactions in the medication of medical patients at hospital discharge. *European journal of clinical pharmacology.* 2003;58(11):773-778.
5. Hansten PD, Horn JR. Hansten and Horn's Drug Interactions Analysis and Management: A Clinical Perspective and Analysis of Current Developments. Applied Therapeutics Incorporated; 2004.
6. Jankel CA, Speedie SM. Detecting drug interactions: a review of the literature. *Dicp.* 1990;24(10):982-989.
7. McInnes GT, Brodie MJ. Drug interactions that matter. *Drugs.* 1988;36(1):83-110.
8. Einarson TR. Drug-related hospital admissions. In: SAGE Publications Sage CA: Los Angeles, CA; 1993.
9. Bjerrum L, Søgaard J, Hallas J, Kragstrup J. Polypharmacy in general practice: differences between practitioners. *British Journal of General Practice.* 1999;49(440):195-198.
10. Reis AMM, Cassiani SHDB. Prevalence of potential drug interactions in patients in an intensive care unit of a university hospital in Brazil. *Clinics.* 2011;66(1):9-15.
11. Moura CS, Acurcio FA, Belo NO. Drug-drug interactions associated with length of stay and cost of hospitalization. *Journal of Pharmacy & Pharmaceutical Sciences.* 2009;12(3):266-272.
12. Shad MU, Marsh C, Preskorn SH. The economic consequences of a drug-drug interaction. *Journal of clinical psychopharmacology.* 2001;21(1):119-120.
13. Bucşa C, Farcaş A, Cazacu I, et al. How many potential drug–drug interactions cause adverse drug reactions in hospitalized patients? *European journal of internal medicine.* 2013;24(1):27-33.
14. Kuperman GJ, Bobb A, Payne TH, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. *Journal of the American Medical Informatics Association.* 2007;14(1):29-40.
15. FOLK1A: Population at the first day of the quarter by region, sex, age and marital status. Statistics Denmark. www.statbank.dk/FOLK1A Accessed 24 March 2021.
16. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clinical epidemiology.* 2015;7:449-490.
17. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: From health care contacts to database records. In. Vol 11: Dove Press; 2019:563-591.
18. Jensen TB, Jimenez-Solem E, Cortes R, et al. Content and validation of the Electronic Patient Medication module (EPM)—the administrative in-hospital drug use database

in the Capital Region of Denmark. *Scandinavian Journal of Public Health.* 2018(October 2017):140349481876005-140349481876005.
19. Oslo N. WHO Collaborating Centre for Drug Statistics Methodology. *ATC classification index with DDDs.* 2017.
20. Aagaard L, Kristensen M. The national drug interactions database. *Ugeskrift for Laeger.* 2005;167(35):3283-3286.
21. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in statistics-simulation and computation.* 2009;38(6):1228-1234.
22. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* Lippincott Williams & Wilkins; 2008.
23. Sloane D, Morgan SP. An introduction to categorical data analysis. *Annual review of sociology.* 1996;22(1):351-375.
24. Walker AM, Patrick AR, Lauer MS, et al. A tool for assessing the feasibility of comparative effectiveness research. *Comp Eff Res.* 2013;2013(3):11-20.
25. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology.* 2006;59(5):437. e431-437. e424.
26. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass).* 2009;20(4):512.
27. Raghunathan K, Layton JB, Ohnuma T, Shaw AD. Observational Research Using Propensity Scores. *Advances in chronic kidney disease.* 2016;23(6):367-372.
28. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Transactions on Modeling and Computer Simulation (TOMACS).* 2013;23(1):1-17.
29. Moore BJ, White S, Washington R, Coenen N, Elixhauser A. Identifying increased risk of readmission and in-hospital mortality using hospital administrative data. *Medical care.* 2017;55(7):698-705.
30. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* 2012;28(19):2520-2522.
31. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Annals of internal medicine.* 2007;147(8):573-577.
32. Percha B, Altman RB. Informatics confronts drug-drug interactions. In. Vol 34: Elsevier Current Trends; 2013:178-184.
33. Hines LE, Murphy JE. Potentially harmful drug–drug interactions in the elderly: a review. *The American journal of geriatric pharmacotherapy.* 2011;9(6):364-377.
34. Chiatti C, Bustacchini S, Furneri G, et al. The economic burden of inappropriate drug prescribing, lack of adherence and compliance, adverse drug events in older people. *Drug safety.* 2012;35(1):73-87.
35. Gnjidic D, Johnell K. Clinical implications from drug-drug and drug-disease interactions in older people. *Clinical and Experimental Pharmacology and Physiology.* 2013;40(5):320-325.

36. Gonzaga de Andrade Santos TN, Mendonça da Cruz Macieira G, Cardoso Sodré Alves BM, et al. Prevalence of clinically manifested drug interactions in hospitalized patients: A systematic review and meta-analysis. *PloS one.* 2020;15(7):e0235353.
37. Muñoz-Torrero JFS, Barquilla P, Velasco R, et al. Adverse drug reactions in internal medicine units and associated risk factors. *European journal of clinical pharmacology.* 2010;66(12):1257-1264.
38. Cruciol-Souza JM, Thomson JC. Prevalence of potential drug-drug interactions and its associated factors in a Brazilian teaching hospital. *J Pharm Pharm Sci.* 2006;9(3):427-433.
39. Koenig Á, Adrieno G. Potential drug interactions prevalence in intensive care units. *Rev Bras Ter Intensiva.* 2008;20(4):349-354.
40. Straubhaar B, Krähenbühl S, Schlienger RG. The prevalence of potential drug-drug interactions in patients with heart failure at hospital discharge. *Drug safety.* 2006;29(1):79-90.
41. Becker ML, Kallewaard M, Caspers PW, Visser LE, Leufkens HG, Stricker BH. Hospitalisations and emergency department visits due to drug–drug interactions: a literature review. *Pharmacoepidemiology and drug safety.* 2007;16(6):641-651.
42. Rosholm J-U, Bjerrum L, Hallas J, Worm J, Gram LF. Polypharmacy and the risk of drug-drug interactions among Danish elderly. A prescription database study. *Danish medical bulletin.* 1998;45(2):210-213.
43. Brattig Correia R, de Araújo Kohler LP, Mattos MM, Rocha LM. City-wide electronic health records reveal gender and age biases in administration of known drug–drug interactions. *npj Digital Medicine.* 2019;2(1):74-74.
44. Glintborg B, Andersen SE, Dalhoff K. Drug-drug interactions among recently hospitalised patients–frequent but mostly clinically insignificant. *European journal of clinical pharmacology.* 2005;61(9):675-681.
45. Gibbs MA, Thummel KE, Shen DD, Kunze KL. Inhibition of cytochrome P-450 3A (CYP3A) in human intestinal and liver microsomes: comparison of K i values and impact of CYP3A5 expression. *Drug Metabolism and Disposition.* 1999;27(2):180-187.

Online supplementary content

Leal and Kaas-Hansen et al., Drug interactions in hospital prescriptions in Denmark: Prevalence and associations with adverse outcomes

eTable 1. Prevalence of potential drug-drug interactions based on recommendation and documentation level.

Unique drug combinations (upper cells) and prevalence (lower cells) of pDDIs by management recommendation and documentation level. Values are N (%).

Recommen-dation level	Documentation level				
	Well documented	Docu-mented	Limited documented	Poorly documented	Total
1: Discouraged	25 (17)	70 (48)	23 (16)	28 (19)	146 (3)
	8,160 (45)	5,918 (33)	3,477 (19)	2,133 (12)	18,192 (3)
2: Dose adjustment	173 (17)	540 (54)	74 (7)	218 (22)	1,005 (18)
	8,1407 (55)	91,096 (61)	3,599 (2)	23,730 (16)	148,455 (23)
3: Staggered ingestion	105 (46)	86 (38)	21 (9)	14 (6)	226 (4)
	81,050 (89)	12,129 (13)	2,744 (3)	488 (1)	90,662 (14)
4: Precautions	304 (17)	834 (48)	161 (9)	443 (25)	1,742 (31)
	16,3097 (31)	468,888 (88)	85,764 (16)	123,792 (23)	532,066 (81)
5: No action needed	259 (10)	1,755 (69)	230 (9)	283 (11)	2,527 (45)
	14,1733 (30)	399,115 (85)	216,539 (46)	170,043 (36)	470,956 (71)
Total	866 (12)	3,285 (58)	509 (9)	986 (17)	5,646
	28,8210 (44)	604,999 (92)	246,771 (37)	234,205 (36)	659,525

eTable 2. Drugs involved in discouraged drug pairs.

Drugs used by less than 5 patients were omitted to avoid privacy issues.

Drug name	ATC	Discouraged-pDDIs (N)	Pairs (N)	Discouraged-pDDIs involving drug (%)	Admissions with discouraged-pDDIs (N)	Admissions with drug (N)	Admissions with discouraged-pDDIs (%)	Patients with discouraged-pDDI (N)	Patients with drug (N)	Patients with discouraged-pDDI (%)	Total admissions (N)	Total patients (N)	Patients overall (%)	Admissions overall (%)
Erythromycin	J01FA01	8	860	0.93	1573	5665	27.77	1461	5001	29.21	2886227	945475	0.53	0.2
Rifabutin	J04AB04	2	224	0.89	25	101	24.75	10	46	21.74	2886227	945475	0	0
Ketoconazole	D01AC08	6	753	0.8	644	3158	20.39	320	1513	21.15	2886227	945475	0.16	0.11
Diclofenac	S01BC03	1	370	0.27	53	313	16.93	13	177	7.34	2886227	945475	0.02	0.01
Clarithromycin	J01FA09	14	1086	1.29	4022	25792	15.59	3338	21008	15.89	2886227	945475	2.22	0.89
Paroxetine	N06AB05	4	832	0.48	1206	8733	13.81	408	2707	15.07	2886227	945475	0.29	0.3
Domperidone	A03FA03	1	1009	0.1	2872	23213	12.37	2028	10571	19.18	2886227	945475	1.12	0.8
Isocarboxazid	N06AF01	4	400	1	90	789	11.41	36	249	14.46	2886227	945475	0.03	0.03
Warfarin	B01AA03	16	1202	1.33	12570	121653	10.33	8791	42101	20.88	2886227	945475	4.45	4.21
Diclofenac	D11AX18	1	289	0.35	18	197	9.14	6	87	6.9	2886227	945475	0.01	0.01
Imatinib	L01XE01	1	491	0.2	139	1543	9.01	44	337	13.06	2886227	945475	0.04	0.05

Etoricoxib	M01AH05	1	216	0.46	9	106	8.49	6	51	11.76	2886227	945475	0.01	0
Sirolimus	L04AA10	3	434	0.69	168	2085	8.06	30	233	12.88	2886227	945475	0.02	0.07
Ketoconazole	J02AB02	2	248	0.81	12	158	7.59	9	75	12	2886227	945475	0.01	0.01
Phenprocoumon	B01AA04	2	542	0.37	146	1985	7.36	97	649	14.95	2886227	945475	0.07	0.07
Pravastatin	C10AA03	1	732	0.14	371	5193	7.14	114	1518	7.51	2886227	945475	0.16	0.18
Ciclosporin	L04AD01	5	819	0.61	681	9595	7.1	193	2257	8.55	2886227	945475	0.24	0.33
Clonidine	N02CX02	3	982	0.31	806	11522	7	303	7371	4.11	2886227	945475	0.78	0.4
Piroxicam	M02AA07	1	401	0.25	27	387	6.98	13	232	5.6	2886227	945475	0.02	0.01
Tizanidine	M03BX02	3	703	0.43	321	4744	6.77	194	1244	15.59	2886227	945475	0.13	0.16
Dronedarone	C01BD07	1	307	0.33	59	910	6.48	32	325	9.85	2886227	945475	0.03	0.03
Tenoxicam	M01AC02	1	271	0.37	15	246	6.1	5	105	4.76	2886227	945475	0.01	0.01
Diclofenac	M02AA15	1	443	0.23	24	433	5.54	16	253	6.32	2886227	945475	0.03	0.02
Voriconazole	J02AC03	3	734	0.41	192	3542	5.42	49	1418	3.46	2886227	945475	0.15	0.12
Itraconazole	J02AC02	2	578	0.35	53	1149	4.61	38	532	7.14	2886227	945475	0.06	0.04
Fluconazole	J02AC01	2	1318	0.15	3802	88426	4.3	2929	50842	5.76	2886227	945475	5.38	3.06
Ibuprofen	M02AA13	1	290	0.34	6	148	4.05	5	94	5.32	2886227	945475	0.01	0.01
Rifampicin, Pyrazinamide, Ethambutol nd Isoniazid	J04AM06	4	303	1.32	12	328	3.66	7	203	3.45	2886227	945475	0.02	0.01

Nabumetone	M01AX01	1	493	0.2	27	759	3.56	18	339	5.31	2886227	945475	0.04	0.03
Simvastatin and Ezetimibe	C10BA02	2	284	0.7	10	293	3.41	8	101	7.92	2886227	945475	0.01	0.01
Terbinafine	D01BA02	5	631	0.79	57	1685	3.38	31	912	3.4	2886227	945475	0.1	0.06
Carbamazepine	N03AF01	8	873	0.92	337	10396	3.24	178	3167	5.62	2886227	945475	0.33	0.36
Terbinafine	D01AE15	5	848	0.59	211	6839	3.09	144	4481	3.21	2886227	945475	0.47	0.24
Diclofenac, combinations	M01AB55	1	546	0.18	37	1371	2.7	22	598	3.68	2886227	945475	0.06	0.05
Etodolac	M01AB08	1	847	0.12	370	14021	2.64	251	9338	2.69	2886227	945475	0.99	0.49
Azathioprine	L04AX01	1	896	0.11	315	12493	2.52	139	3640	3.82	2886227	945475	0.38	0.43
Ritonavir	J05AE03	5	590	0.85	76	3029	2.51	47	792	5.93	2886227	945475	0.08	0.1
Amitriptyline	N06AA09	3	1114	0.27	902	36127	2.5	376	12365	3.04	2886227	945475	1.31	1.25
Emtricitabine, tenofovir disoproxil, elvitegravir and cobicistat	J05AR09	4	232	1.72	5	209	2.39	5	93	5.38	2886227	945475	0.01	0.01
Fluoxetine	N06AB03	2	875	0.23	215	10164	2.12	150	3450	4.35	2886227	945475	0.36	0.35
Colchicine	M04AC01	1	779	0.13	165	8144	2.03	93	3728	2.49	2886227	945475	0.39	0.28
Agomelatine	N06AX22	2	672	0.3	76	3810	1.99	57	1596	3.57	2886227	945475	0.17	0.13
Tacrolimus	D11AH01	2	458	0.44	11	562	1.96	10	252	3.97	2886227	945475	0.03	0.02
Posaconazole	J02AC04	1	628	0.16	90	4768	1.89	38	964	3.94	2886227	945475	0.1	0.17

Rifampicin	J04AB02	6	857	0.7	131	7229	1.81	102	4308	2.37	2886227	945475	0.46	0.25
Valproic acid	N03AG01	2	1029	0.19	606	33631	1.8	458	10000	4.58	2886227	945475	1.06	1.17
Ibuprofen	M01AE01	2	1437	0.14	9982	569223	1.75	7368	365302	2.02	2886227	945475	38.64	19.72
Clomipramine	N06AA04	3	596	0.5	48	2852	1.68	21	824	2.55	2886227	945475	0.09	0.1
Drospirenone and ethinylestradiol	G03AA12	3	419	0.72	20	1220	1.64	7	412	1.7	2886227	945475	0.04	0.04
Desogestrel	G03AC09	4	335	1.19	7	436	1.61	5	201	2.49	2886227	945475	0.02	0.02
Medroxyprogesterone	G03DA02	4	488	0.82	17	1156	1.47	5	374	1.34	2886227	945475	0.04	0.04
Ertapenem	J01DH03	1	697	0.14	41	3009	1.36	32	2076	1.54	2886227	945475	0.22	0.1
Estriol	G03CA04	3	645	0.47	39	3075	1.27	14	1211	1.16	2886227	945475	0.13	0.11
Oxcarbazepine	N03AF02	4	868	0.46	131	10597	1.24	52	2986	1.74	2886227	945475	0.32	0.37
Methenamine	J01XX05	1	620	0.16	31	2563	1.21	27	906	2.98	2886227	945475	0.1	0.09
Atazanavir	J05AE08	3	507	0.59	22	1894	1.16	12	566	2.12	2886227	945475	0.06	0.07
Simvastatin	C10AA01	8	1392	0.57	5048	442545	1.14	3887	148579	2.62	2886227	945475	15.71	15.33
Estradiol	G03CA03	5	1058	0.47	344	31379	1.1	138	10795	1.28	2886227	945475	1.14	1.09
Diclofenac	M01AB05	1	1245	0.08	1917	177928	1.08	1326	120256	1.1	2886227	945475	12.72	6.16
Meropenem	J01DH02	1	1283	0.08	575	55688	1.03	439	36151	1.21	2886227	945475	3.82	1.93
Imipramine	N06AA02	3	785	0.38	69	6844	1.01	45	2235	2.01	2886227	945475	0.24	0.24

Tacrolimus	L04AD02	2	765	0.26	90	9708	0.93	68	1900	3.58	2886227	945475	0.2	0.34
Celecoxib	M01AH01	1	996	0.1	661	72096	0.92	571	57153	1	2886227	945475	6.04	2.5
Naproxen	M01AE02	2	811	0.25	165	18264	0.9	110	12381	0.89	2886227	945475	1.31	0.63
Everolimus	L04AA18	2	459	0.44	15	1735	0.86	12	294	4.08	2886227	945475	0.03	0.06
Selegiline	N04BD01	2	494	0.4	14	1698	0.82	6	544	1.1	2886227	945475	0.06	0.06
Venlafaxine	N06AX16	3	1082	0.28	220	39070	0.56	149	12541	1.19	2886227	945475	1.33	1.35
Verapamil	C08DA01	3	1036	0.29	221	41236	0.54	135	13786	0.98	2886227	945475	1.46	1.43
Tamoxifen	L02BA01	2	620	0.32	21	4280	0.49	16	1974	0.81	2886227	945475	0.21	0.15
Phenobarbital	N03AA02	9	945	0.95	107	24658	0.43	51	10612	0.48	2886227	945475	1.12	0.85
Allopurinol	M04AA01	1	1181	0.08	315	76643	0.41	139	22843	0.61	2886227	945475	2.42	2.66
Quetiapine	N05AH04	1	1134	0.09	297	74846	0.4	234	24677	0.95	2886227	945475	2.61	2.59
Nortriptyline	N06AA10	2	1045	0.19	87	21962	0.4	60	7232	0.83	2886227	945475	0.76	0.76
Ticagrelor	B01AC24	2	984	0.2	188	49328	0.38	155	30626	0.51	2886227	945475	3.24	1.71
Mianserin	N06AX03	1	930	0.11	59	16501	0.36	21	6006	0.35	2886227	945475	0.64	0.57
Probenecid	M04AB01	2	551	0.36	5	1435	0.35	5	469	1.07	2886227	945475	0.05	0.05
Midazolam	N05CD08	2	1137	0.18	145	42109	0.34	108	33001	0.33	2886227	945475	3.49	1.46
Metoprolol	C07AB02	1	1392	0.07	1191	379785	0.31	399	127237	0.31	2886227	945475	13.46	13.16

Apixaban	B01AF02	1	811	0.12	49	16377	0.3	24	8230	0.29	2886227	945475	0.87	0.57
Nifedipine	C08CA05	2	920	0.22	36	12454	0.29	30	4457	0.67	2886227	945475	0.47	0.43
Ciprofloxacin	J01MA02	2	1384	0.14	393	135256	0.29	248	86572	0.29	2886227	945475	9.16	4.69
Letrozole	L02BG04	1	761	0.13	25	9502	0.26	14	4067	0.34	2886227	945475	0.43	0.33
Sulfamethizole	J01EB02	1	943	0.11	31	11944	0.26	27	10010	0.27	2886227	945475	1.06	0.41
Dabigatran etexilate	B01AE07	2	899	0.22	81	32015	0.25	43	14549	0.3	2886227	945475	1.54	1.11
Atorvastatin	C10AA05	1	1142	0.09	231	94086	0.25	64	36004	0.18	2886227	945475	3.81	3.26
Rosuvastatin	C10AA07	1	967	0.1	54	23416	0.23	10	7811	0.13	2886227	945475	0.83	0.81
Zuclopenthixol	N05AF05	1	823	0.12	25	14838	0.17	14	4039	0.35	2886227	945475	0.43	0.51
Cabergoline	G02CB03	1	506	0.2	10	5821	0.17	10	4909	0.2	2886227	945475	0.52	0.2
Triazolam	N05CD05	3	1118	0.27	43	66578	0.06	33	48423	0.07	2886227	945475	5.12	2.31
Mirtazapine	N06AX11	1	1203	0.08	28	76274	0.04	13	27409	0.05	2886227	945475	2.9	2.64
Rivaroxaban	B01AF01	1	973	0.1	9	53062	0.02	5	34144	0.01	2886227	945475	3.61	1.84
Lansoprazole	A02BC03	1	1243	0.08	15	118737	0.01	6	43814	0.01	2886227	945475	4.63	4.11
Pantoprazole	A02BC02	2	1518	0.13	9	570440	0	5	224002	0	2886227	945475	23.69	19.76

eTable 3. Overview of drugs involved in discouraged combinations

Drugs used by less than 5 patients were omitted to avoid privacy issues.

Drug name	ATC	Recommendation level	Clinical significance	Documentation	pDDIs (N)
Domperidone	A03FA03	5	Minor	Documented	2
Domperidone	A03FA03	5	None	Documented	2
Domperidone	A03FA03	1	Major	Limited documented	1
Warfarin	B01AA03	5	Minor	Documented	3
Warfarin	B01AA03	5	Minor	Well documented	3
Warfarin	B01AA03	5	Moderate	Poorly documented	1
Warfarin	B01AA03	5	None	Documented	56
Warfarin	B01AA03	5	None	Limited documented	4
Warfarin	B01AA03	5	None	Well documented	4
Warfarin	B01AA03	5	Possible	Documented	4
Warfarin	B01AA03	5	Undetermined	Documented	4
Warfarin	B01AA03	5	Undetermined	Poorly documented	12
Warfarin	B01AA03	5	Undetermined	Well documented	2
Warfarin	B01AA03	4	Major	Well documented	5
Warfarin	B01AA03	4	Minor	Documented	1
Warfarin	B01AA03	4	Moderate	Documented	6
Warfarin	B01AA03	4	Moderate	Limited documented	4
Warfarin	B01AA03	4	Moderate	Poorly documented	5
Warfarin	B01AA03	4	Moderate	Well documented	17
Warfarin	B01AA03	4	Possible	Documented	11
Warfarin	B01AA03	4	Possible	Limited documented	3
Warfarin	B01AA03	4	Possible	Poorly documented	5
Warfarin	B01AA03	4	Possible	Well documented	1
Warfarin	B01AA03	4	Undetermined	Documented	2
Warfarin	B01AA03	4	Undetermined	Poorly documented	3
Warfarin	B01AA03	2	Major	Documented	4
Warfarin	B01AA03	2	Major	Limited documented	4

Warfarin	B01AA03	2	Major	Well documented	2
Warfarin	B01AA03	2	Moderate	Documented	13
Warfarin	B01AA03	2	Moderate	Limited documented	1
Warfarin	B01AA03	2	Moderate	Poorly documented	3
Warfarin	B01AA03	2	Moderate	Well documented	6
Warfarin	B01AA03	2	Possible	Documented	1
Warfarin	B01AA03	2	Possible	Poorly documented	6
Warfarin	B01AA03	1	Major	Documented	5
Warfarin	B01AA03	1	Major	Poorly documented	6
Warfarin	B01AA03	1	Major	Well documented	4
Warfarin	B01AA03	1	Possible	Documented	1
Phenprocoumon	B01AA04	5	None	Documented	12
Phenprocoumon	B01AA04	5	None	Limited documented	2
Phenprocoumon	B01AA04	5	None	Poorly documented	2
Phenprocoumon	B01AA04	5	None	Well documented	1
Phenprocoumon	B01AA04	5	Undetermined	Limited documented	2
Phenprocoumon	B01AA04	4	Moderate	Documented	4
Phenprocoumon	B01AA04	4	Moderate	Poorly documented	4
Phenprocoumon	B01AA04	4	Possible	Documented	7
Phenprocoumon	B01AA04	4	Possible	Limited documented	2
Phenprocoumon	B01AA04	4	Possible	Poorly documented	2
Phenprocoumon	B01AA04	4	Undetermined	Limited documented	1
Phenprocoumon	B01AA04	4	Undetermined	Poorly documented	1
Phenprocoumon	B01AA04	2	Moderate	Documented	1
Phenprocoumon	B01AA04	2	Moderate	Poorly documented	2
Phenprocoumon	B01AA04	2	Possible	Poorly documented	1
Phenprocoumon	B01AA04	2	Undetermined	Limited documented	1
Phenprocoumon	B01AA04	1	Major	Documented	1
Phenprocoumon	B01AA04	1	Major	Well documented	1
Ketoconazole	D01AC08	5	Minor	Documented	3
Ketoconazole	D01AC08	5	Minor	Limited documented	3

Ketoconazole	D01AC08	5	None	Documented	11
Ketoconazole	D01AC08	5	None	Limited documented	5
Ketoconazole	D01AC08	5	Possible	Well documented	1
Ketoconazole	D01AC08	5	Undetermined	Documented	1
Ketoconazole	D01AC08	4	Major	Documented	2
Ketoconazole	D01AC08	4	Moderate	Documented	5
Ketoconazole	D01AC08	4	Possible	Documented	11
Ketoconazole	D01AC08	4	Possible	Well documented	1
Ketoconazole	D01AC08	3	Major	Documented	1
Ketoconazole	D01AC08	3	Possible	Poorly documented	7
Ketoconazole	D01AC08	2	Major	Documented	3
Ketoconazole	D01AC08	2	Major	Well documented	3
Ketoconazole	D01AC08	2	Minor	Documented	1
Ketoconazole	D01AC08	2	Moderate	Documented	20
Ketoconazole	D01AC08	2	Moderate	Poorly documented	1
Ketoconazole	D01AC08	2	Moderate	Well documented	3
Ketoconazole	D01AC08	2	Possible	Documented	1
Ketoconazole	D01AC08	1	Major	Documented	3
Ketoconazole	D01AC08	1	Major	Limited documented	3
Erythromycin	J01FA01	5	Minor	Documented	3
Erythromycin	J01FA01	5	Moderate	Poorly documented	1
Erythromycin	J01FA01	5	None	Documented	7
Erythromycin	J01FA01	5	None	Limited documented	1
Erythromycin	J01FA01	4	Major	Documented	2
Erythromycin	J01FA01	4	Major	Poorly documented	2
Erythromycin	J01FA01	4	Major	Well documented	1
Erythromycin	J01FA01	4	Moderate	Documented	2
Erythromycin	J01FA01	4	Possible	Documented	3
Erythromycin	J01FA01	4	Possible	Limited documented	2
Erythromycin	J01FA01	4	Possible	Well documented	1
Erythromycin	J01FA01	4	Undetermined	Documented	6
Erythromycin	J01FA01	4	Undetermined	Limited documented	1
Erythromycin	J01FA01	4	Undetermined	Poorly documented	1

Erythromycin	J01FA01	2	Major	Documented	1
Erythromycin	J01FA01	2	Major	Limited documented	1
Erythromycin	J01FA01	2	Moderate	Documented	6
Erythromycin	J01FA01	2	Moderate	Well documented	1
Erythromycin	J01FA01	2	Possible	Well documented	1
Erythromycin	J01FA01	2	Undetermined	Limited documented	1
Erythromycin	J01FA01	1	Major	Documented	2
Erythromycin	J01FA01	1	Major	Limited documented	3
Erythromycin	J01FA01	1	Major	Well documented	2
Erythromycin	J01FA01	1	Possible	Documented	1
Clarithromycin	J01FA09	5	Minor	Documented	4
Clarithromycin	J01FA09	5	Minor	Limited documented	1
Clarithromycin	J01FA09	5	Moderate	Poorly documented	1
Clarithromycin	J01FA09	5	None	Documented	4
Clarithromycin	J01FA09	5	None	Limited documented	1
Clarithromycin	J01FA09	5	Possible	Documented	1
Clarithromycin	J01FA09	5	Possible	Limited documented	1
Clarithromycin	J01FA09	5	Possible	Poorly documented	2
Clarithromycin	J01FA09	5	Undetermined	Limited documented	1
Clarithromycin	J01FA09	5	Undetermined	Poorly documented	2
Clarithromycin	J01FA09	4	Major	Documented	3
Clarithromycin	J01FA09	4	Major	Poorly documented	14
Clarithromycin	J01FA09	4	Moderate	Documented	1
Clarithromycin	J01FA09	4	Moderate	Limited documented	2
Clarithromycin	J01FA09	4	Moderate	Poorly documented	1
Clarithromycin	J01FA09	4	Moderate	Well documented	2
Clarithromycin	J01FA09	4	None	Documented	1
Clarithromycin	J01FA09	4	Possible	Documented	7
Clarithromycin	J01FA09	4	Possible	Limited documented	2
Clarithromycin	J01FA09	4	Possible	Well documented	2
Clarithromycin	J01FA09	4	Undetermined	Poorly documented	2

Clarithromycin	J01FA09	3	Moderate	Poorly documented	2
Clarithromycin	J01FA09	2	Major	Well documented	1
Clarithromycin	J01FA09	2	Moderate	Documented	6
Clarithromycin	J01FA09	2	Moderate	Limited documented	1
Clarithromycin	J01FA09	2	Moderate	Well documented	1
Clarithromycin	J01FA09	2	Possible	Documented	2
Clarithromycin	J01FA09	2	Possible	Limited documented	2
Clarithromycin	J01FA09	2	Possible	Well documented	1
Clarithromycin	J01FA09	2	Undetermined	Documented	2
Clarithromycin	J01FA09	1	Major	Documented	8
Clarithromycin	J01FA09	1	Major	Poorly documented	4
Clarithromycin	J01FA09	1	Moderate	Documented	2
Ketoconazole	J02AB02	5	None	Documented	6
Ketoconazole	J02AB02	5	Undetermined	Documented	1
Ketoconazole	J02AB02	4	Moderate	Documented	2
Ketoconazole	J02AB02	4	Possible	Documented	1
Ketoconazole	J02AB02	3	Possible	Poorly documented	4
Ketoconazole	J02AB02	2	Major	Documented	1
Ketoconazole	J02AB02	2	Major	Well documented	2
Ketoconazole	J02AB02	2	Moderate	Documented	5
Ketoconazole	J02AB02	2	Moderate	Poorly documented	1
Ketoconazole	J02AB02	2	Moderate	Well documented	3
Ketoconazole	J02AB02	1	Major	Documented	1
Ketoconazole	J02AB02	1	Major	Limited documented	1
Rifabutin	J04AB04	5	Minor	Documented	1
Rifabutin	J04AB04	5	Minor	Poorly documented	1
Rifabutin	J04AB04	5	None	Documented	3
Rifabutin	J04AB04	4	Moderate	Documented	1
Rifabutin	J04AB04	4	Moderate	Well documented	1
Rifabutin	J04AB04	4	Possible	Documented	1
Rifabutin	J04AB04	2	Major	Documented	1
Rifabutin	J04AB04	2	Moderate	Documented	3
Rifabutin	J04AB04	2	Possible	Well documented	1
Rifabutin	J04AB04	1	Major	Documented	1

Rifabutin	J04AB04	1	Moderate	Documented	1
Imatinib	L01XE01	5	None	Documented	9
Imatinib	L01XE01	5	Possible	Poorly documented	1
Imatinib	L01XE01	5	Undetermined	Poorly documented	1
Imatinib	L01XE01	4	Minor	Documented	1
Imatinib	L01XE01	4	Moderate	Documented	1
Imatinib	L01XE01	4	Moderate	Limited documented	1
Imatinib	L01XE01	4	Moderate	Poorly documented	2
Imatinib	L01XE01	4	Possible	Limited documented	1
Imatinib	L01XE01	4	Possible	Poorly documented	1
Imatinib	L01XE01	2	Moderate	Documented	1
Imatinib	L01XE01	2	Possible	Documented	2
Imatinib	L01XE01	1	Major	Documented	1
Sirolimus	L04AA10	5	None	Documented	2
Sirolimus	L04AA10	4	Major	Poorly documented	1
Sirolimus	L04AA10	4	Moderate	Limited documented	1
Sirolimus	L04AA10	4	Moderate	Poorly documented	1
Sirolimus	L04AA10	2	Major	Documented	1
Sirolimus	L04AA10	2	Major	Limited documented	3
Sirolimus	L04AA10	2	Moderate	Limited documented	1
Sirolimus	L04AA10	2	Possible	Well documented	1
Sirolimus	L04AA10	2	Undetermined	Documented	1
Sirolimus	L04AA10	1	Major	Documented	1
Sirolimus	L04AA10	1	Major	Limited documented	1
Sirolimus	L04AA10	1	Major	Poorly documented	1
Etoricoxib	M01AH05	5	Minor	Documented	2
Etoricoxib	M01AH05	5	None	Documented	3
Etoricoxib	M01AH05	4	Undetermined	Poorly documented	1
Etoricoxib	M01AH05	2	Possible	Documented	1
Etoricoxib	M01AH05	1	Major	Documented	1
Tizanidine	M03BX02	4	Moderate	Limited documented	3
Tizanidine	M03BX02	4	Possible	Documented	1
Tizanidine	M03BX02	4	Possible	Poorly documented	2

Tizanidine	M03BX02	1	Major	Documented	3
Paroxetine	N06AB05	5	None	Documented	10
Paroxetine	N06AB05	5	Possible	Poorly documented	2
Paroxetine	N06AB05	4	Major	Poorly documented	2
Paroxetine	N06AB05	4	Major	Well documented	1
Paroxetine	N06AB05	4	Minor	Documented	2
Paroxetine	N06AB05	4	Moderate	Documented	3
Paroxetine	N06AB05	4	Moderate	Limited documented	1
Paroxetine	N06AB05	4	Moderate	Poorly documented	11
Paroxetine	N06AB05	4	Moderate	Well documented	2
Paroxetine	N06AB05	4	Possible	Documented	7
Paroxetine	N06AB05	4	Possible	Limited documented	2
Paroxetine	N06AB05	4	Possible	Poorly documented	1
Paroxetine	N06AB05	4	Undetermined	Documented	1
Paroxetine	N06AB05	4	Undetermined	Poorly documented	1
Paroxetine	N06AB05	2	Moderate	Documented	1
Paroxetine	N06AB05	2	None	Documented	1
Paroxetine	N06AB05	1	Major	Documented	3
Paroxetine	N06AB05	1	Moderate	Well documented	1
Isocarboxazid	N06AF01	4	Major	Limited documented	2
Isocarboxazid	N06AF01	4	Major	Poorly documented	1
Isocarboxazid	N06AF01	4	Moderate	Limited documented	3
Isocarboxazid	N06AF01	4	Moderate	Poorly documented	1
Isocarboxazid	N06AF01	4	Possible	Poorly documented	1
Isocarboxazid	N06AF01	1	Major	Poorly documented	3
Isocarboxazid	N06AF01	1	Undetermined	Poorly documented	1

eTable 4. List of discouraged drug pairs

Discouraged drug pairs dispensed in at least 5 admissions (≥5 patients) prescribed during patient hospitalisation.

Drug 1	Drug 2	Admissions (N)
Warfarin (B01AA03)	Ibuprofen (M01AE01)	5,960
Simvastatin (C10AA01)	Clarithromycin (J01FA09)	2,354
Domperidone (A03FA03)	Fluconazole (J02AC01)	1,664
Erythromycin (J01FA01)	Fluconazole (J02AC01)	917
Warfarin (B01AA03)	Diclofenac (M01AB05)	815
Simvastatin (C10AA01)	Erythromycin (J01FA01)	525
Meropenem (J01DH02)	Valproic acid (N03AG01)	444
Warfarin (B01AA03)	Celecoxib (M01AH01)	327
Ciprofloxacin (J01MA02)	Tizanidine (M03BX02)	204
Clarithromycin (J01FA09)	Quetiapine (N05AH04)	194
Metoprolol (C07AB02)	Paroxetine (N06AB05)	175
Simvastatin (C10AA01)	Ketoconazole (D01AC08)	173
Clonidine (N02CX02)	Amitriptyline (N06AA09)	163
Warfarin (B01AA03)	Etodolac (M01AB08)	119
Ticagrelor (B01AC24)	Clarithromycin (J01FA09)	93
Terbinafine (D01AE15)	Amitriptyline (N06AA09)	84
Fluoxetine (N06AB03)	Venlafaxine (N06AX16)	81
Phenprocoumon (B01AA04)	Ibuprofen (M01AE01)	76
Carbamazepine (N03AF01)	Midazolam (N05CD08)	74
Pravastatin (C10AA03)	Ciclosporin (L04AD01)	72
Verapamil (C08DA01)	Colchicine (M04AC01)	70
Warfarin (B01AA03)	Naproxen (M01AE02)	64
Azathioprine (L04AX01)	Allopurinol (M04AA01)	58
Ciprofloxacin (J01MA02)	Agomelatine (N06AX22)	45
Terbinafine (D01AE15)	Nortriptyline (N06AA10)	43
Simvastatin (C10AA01)	Posaconazole (J02AC04)	42
Atorvastatin (C10AA05)	Ciclosporin (L04AD01)	39
Erythromycin (J01FA01)	Tacrolimus (L04AD02)	38
Verapamil (C08DA01)	Rifampicin (J04AB02)	29
Clarithromycin (J01FA09)	Tacrolimus (L04AD02)	27
Estradiol (G03CA03)	Carbamazepine (N03AF01)	26
Nifedipine (C08CA05)	Rifampicin (J04AB02)	24
Voriconazole (J02AC03)	Rifampicin (J04AB02)	22
Ketoconazole (D01AC08)	Clarithromycin (J01FA09)	21

Ertapenem (J01DH03)	Valproic acid (N03AG01)	20
Sulfamethizole (J01EB02)	Methenamine (J01XX05)	19
Ritonavir (J05AE03)	Triazolam (N05CD05)	19
Simvastatin (C10AA01)	Imatinib (L01XE01)	19
Simvastatin (C10AA01)	Itraconazole (J02AC02)	18
Voriconazole (J02AC03)	Sirolimus (L04AA10)	18
Dabigatran etexilate (B01AE07)	Dronedarone (C01BD07)	17
Estradiol (G03CA03)	Phenobarbital (N03AA02)	16
Clonidine (N02CX02)	Imipramine (N06AA02)	15
Apixaban (B01AF02)	Ketoconazole (D01AC08)	15
Isocarboxazid (N06AF01)	Mianserin (N06AX03)	14
Terbinafine (D01AE15)	Imipramine (N06AA02)	14
Estradiol (G03CA03)	Oxcarbazepine (N03AF02)	12
Warfarin (B01AA03)	Diclofenac, combinations (M01AB55)	12
Warfarin (B01AA03)	Piroxicam (M02AA07)	11
Warfarin (B01AA03)	Diclofenac (M02AA15)	11
Terbinafine (D01BA02)	Amitriptyline (N06AA09)	10
Warfarin (B01AA03)	Nabumetone (M01AX01)	10
Clarithromycin (J01FA09)	Everolimus (L04AA18)	9
Letrozole (L02BG04)	Zuclopenthixol (N05AF05)	8
Warfarin (B01AA03)	Diclofenac (S01BC03)	8
Terbinafine (D01BA02)	Nortriptyline (N06AA10)	7
Rifabutin (J04AB04)	Ritonavir (J05AE03)	7
Isocarboxazid (N06AF01)	Mirtazapine (N06AX11)	7
Clonidine (N02CX02)	Clomipramine (N06AA04)	7
Ticagrelor (B01AC24)	Ciclosporin (L04AD01)	7
Rifampicin (J04AB02)	Atazanavir (J05AE08)	5
Cabergoline (G02CB03)	Clarithromycin (J01FA09)	5
Dabigatran etexilate (B01AE07)	Ketoconazole (D01AC08)	5
Tacrolimus (D11AH01)	Clarithromycin (J01FA09)	5
Lansoprazole (A02BC03)	Atazanavir (J05AE08)	5

eTable 5. Effect-size estimates of exposure to discouraged drug pairs on post-discharge mortality, readmission and length-of-stay

Provides numerical estimate illustrated in Figure 3: estimate (95 confidence interval). Mortality and readmission estimates are hazard ratios. Values below 1 are shown with 2 significant digits, otherwise with 1 significant digit.

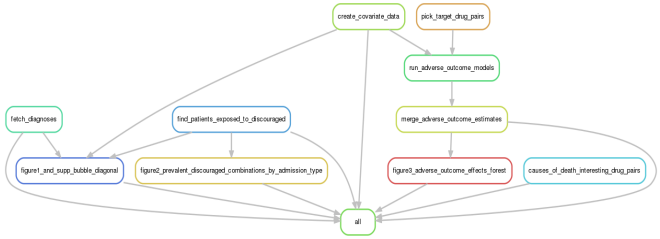
Clinical significance	Drug pair	Mortality	Readmission	LOS
Moderate	Estradiol + Phenobarbital	-	0.74 (0.16; 3.4)	0.21 (-0.088; 0.51)
Moderate	Carbamazepine + Estradiol	0.45 (0.1; 2)	0.97 (0.47; 2)	0.2 (0.0049; 0.39)
Moderate	Estradiol + Oxcarbazepine	0.78 (0.21; 2.9)	2.2 (0.81; 6)	-0.036 (-0.23; 0.15)
Major	Naproxen + Warfarin	1.3 (0.61; 2.7)	0.92 (0.56; 1.5)	-0.14 (-0.24; -0.032)
Major	Apixaban + Ketoconazole	2.5 (0.81; 7.7)	1.6 (0.82; 3.2)	-0.022 (-0.18; 0.14)
Major	Erythromycin + Fluconazole	0.99 (0.73; 1.3)	0.72 (0.58; 0.89)	-0.089 (-0.12; -0.056)
Major	Imatinib + Simvastatin	1.1 (0.4; 3.2)	1.5 (0.66; 3.6)	-0.12 (-0.38; 0.13)
Major	Atorvastatin + Ciclosporin	0.73 (0.24; 2.3)	1.7 (0.78; 3.6)	0.77 (0.64; 0.9)
Major	Carbamazepine + Midazolam	1.5 (0.77; 2.9)	0.7 (0.44; 1.1)	-0.004 (-0.083; 0.075)
Possible	Celecoxib + Warfarin	0.78 (0.45; 1.4)	1.2 (0.85; 1.7)	0.036 (-0.04; 0.11)
Un-determined	Isocarboxazid + Mianserin	-	0.78 (0.15; 3.9)	0.26 (-0.18; 0.7)
Major	Diclofenac (topical) + Warfarin	0.96 (0.26; 3.5)	1.7 (0.66; 4.6)	0.8 (0.64; 0.97)
Major	Posaconazole + Simvastatin	7.9 (0.71; 88)	1.3 (0.27; 6.7)	1.1 (0.91; 1.3)
Major	Ciclosporin + Ticagrelor	-	4.2 (0.58; 29.9)	0.31 (-0.053; 0.66)
Major	Clarithromycin + Tacrolimus (immunosupp.)	2.3 (0.64; 8.4)	3.4 (1.1; 10.4)	0.12 (-0.098; 0.34)

Major	Paroxetine + Tamoxifen	1.4 (0.27; 7.3)	1.9 (0.33; 10.4)	1.1 (0.78; 1.4)
Moderate	Itraconazole + Simvastatin	2.4 (0.75; 7.9)	2.5 (0.92; 6.6)	1.1 (0.93; 1.2)
Major	Diclofenac, combinations + Warfarin	1.5 (0.39; 5.9)	1.6 (0.61; 4.3)	-0.1 (-0.38; 0.17)
Major	Clonidine + Imipramine	0.87 (0.095; 7.9)	0.83 (0.31; 2.3)	0.38 (0.21; 0.54)
Moderate	Meropenem + Valproic acid	1.5 (1.1; 1.9)	0.92 (0.76; 1.1)	0.22 (0.19; 0.26)
Major	Nifedipine + Rifampicin	1.1 (0.13; 10.2)	1.2 (0.45; 3)	0.1 (-0.11; 0.31)
Moderate	Colchicine + Verapamil	0.61 (0.29; 1.3)	1 (0.63; 1.7)	0.13 (0.032; 0.23)
Major	Ketoconazole + Simvastatin	1 (0.68; 1.6)	1.3 (0.91; 1.8)	1.1 (1.1; 1.1)
Moderate	Fluoxetine + Venlafaxine	-	1.3 (0.4; 4.2)	-0.38 (-0.7; -0.063)
Major	Etoricoxib + Warfarin	2.8 (0.17; 47.1)	0.82 (0.17; 3.8)	-0.77 (-1.2; -0.36)
Major	Amitriptyline + Terbinafine (systemic)	2.1 (0.53; 8)	3.1 (1.1; 8.8)	-0.39 (-0.63; -0.15)
Major	Imipramine + Terbinafine	3.8 (1.2; 12)	1 (0.42; 2.4)	0.46 (0.29; 0.62)
Major	Clarithromycin + Everolimus	0.97 (0.1; 9)	12.3 (1.1; 138.7)	-1.2 (-1.9; -0.44)
Major	Ibuprofen + Phenprocoumon	0.81 (0.48; 1.4)	1.2 (0.86; 1.7)	0.2 (0.12; 0.27)
Major	Clarithromycin + Quetiapine	1.7 (1.1; 2.7)	0.94 (0.66; 1.3)	1.6 (1.5; 1.6)
Major	Ritonavir + Triazolam	3.6 (0.49; 26.1)	2.6 (0.84; 8.1)	-0.26 (-0.53; 0.0016)
Major	Rifampicin + Voriconazole	0.65 (0.077; 5.4)	1.3 (0.39; 4.2)	0.067 (-0.13; 0.26)
Major	Piroxicam + Warfarin	3.4 (1; 11.4)	1.4 (0.53; 3.9)	0.034 (-0.19; 0.25)

Major	Diclofenac (systemic) + Warfarin	0.83 (0.69; 1)	1.1 (0.96; 1.2)	0.25 (0.23; 0.28)
Major	Ertapenem + Valproic acid	1.3 (0.54; 3.1)	2.6 (1.1; 5.8)	0.24 (0.11; 0.38)
Major	Dabigatran etexilate + Ketoconazole	-	0.68 (0.14; 3.2)	0.06 (-0.26; 0.38)
Major	Diclofenac (ophthalmologica) + Warfarin	0.83 (0.23; 3)	1.7 (0.72; 4)	0.45 (0.27; 0.62)
Major	Clarithromycin + Ticagrelor	0.98 (0.51; 1.9)	1.7 (1.1; 2.5)	-0.25 (-0.34; -0.16)
Major	Clarithromycin + Ketoconazole	1.2 (0.24; 5.8)	0.56 (0.19; 1.7)	0.096 (-0.12; 0.31)
Moderate	Estriol + Phenobarbital	-	0.45 (0.052; 4)	-0.13 (-0.63; 0.38)
Major	Sirolimus + Voriconazole	0.51 (0.056; 4.6)	2.7 (0.45; 16.5)	-0.21 (-0.74; 0.32)
Major	Ciprofloxacin + Tizanidine	1.4 (0.92; 2.1)	1.4 (0.99; 1.9)	0.36 (0.29; 0.42)
Major	Ciclosporin + Pravastatin	0.22 (0.027; 1.7)	0.86 (0.22; 3.3)	-0.42 (-0.77; -0.073)
Major	Allopurinol + Azathioprine	0.93 (0.39; 2.2)	0.87 (0.49; 1.6)	0.092 (-0.036; 0.22)
Major	Rifabutin + Ritonavir	-	-	0.51 (0.091; 0.93)
Major	Nabumetone + Warfarin	0.99 (0.2; 4.8)	0.67 (0.18; 2.4)	-0.69 (-0.98; -0.39)
Major	Domperidone + Fluconazole	2.5 (2.1; 3.1)	1 (0.89; 1.2)	0.29 (0.27; 0.32)
Major	Methenamine + Sulfamethizole	1.7 (0.64; 4.6)	1.4 (0.67; 2.8)	0.81 (0.7; 0.92)
Major	Erythromycin + Tacrolimus	0.55 (0.066; 4.5)	0.81 (0.22; 3)	0.42 (0.17; 0.67)
Major	Rifampicin + Verapamil	0.7 (0.28; 1.8)	1.4 (0.67; 2.7)	0.36 (0.24; 0.47)
Moderate	Metoprolol + Paroxetine	1.1 (0.72; 1.7)	0.92 (0.67; 1.3)	0.088 (0.026; 0.15)

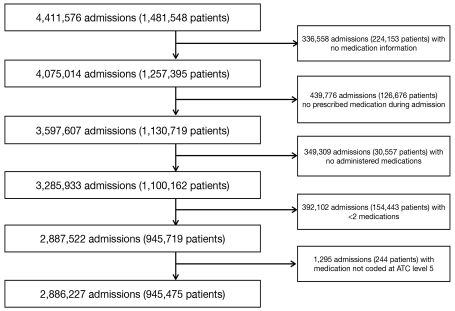
Major	Amitriptyline + Clonidine	1.1 (0.62; 1.8)	1.1 (0.77; 1.6)	-0.16 (-0.23; -0.085)
Major	Ibuprofen + Warfarin	0.9 (0.83; 0.98)	1.1 (1; 1.2)	0.006 (-0.0062; 0.018)
Major	Amitriptyline + Terbinafine (topical)	1.1 (0.66; 1.9)	1 (0.73; 1.5)	0.54 (0.48; 0.6)
Major	Clarithromycin + Simvastatin	1.1 (0.95; 1.2)	1 (0.92; 1.1)	0.082 (0.063; 0.1)
Major	Agomelatine + Ciprofloxacin	2.6 (1.3; 5.5)	1.1 (0.62; 1.9)	0.82 (0.73; 0.9)
Major	Nortriptyline + Terbinafine (systemic)	4.3 (0.86; 21.8)	1.7 (0.43; 6.6)	-0.29 (-0.61; 0.032)
Major	Nortriptyline + Terbinafine (topical)	1.1 (0.46; 2.6)	1.1 (0.65; 2)	0.67 (0.58; 0.77)
Major	Etodolac + Warfarin	0.76 (0.43; 1.3)	1.1 (0.78; 1.6)	-0.15 (-0.23; -0.07)
Possible	Erythromycin + Simvastatin	1.1 (0.83; 1.3)	1.1 (0.93; 1.4)	0.067 (0.033; 0.1)
Major	Clarithromycin + Tacrolimus (topical)	-	0.51 (0.062; 4.2)	-0.087 (-0.54; 0.37)
Major	Clomipramine + Clonidine	0.55 (0.065; 4.6)	1.4 (0.43; 4.8)	-0.22 (-0.47; 0.029)

eFigure 1. Pipeline workflow

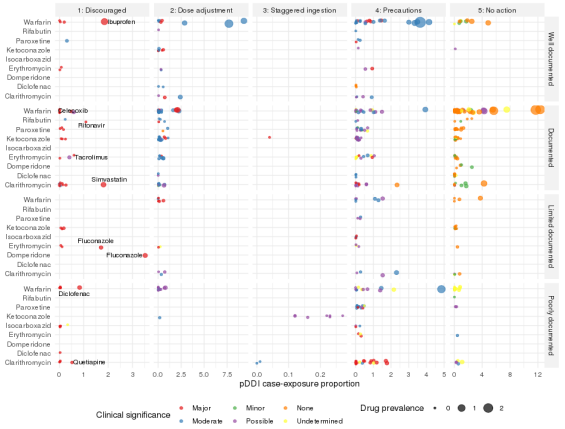


Snakemake pipeline depicting the workflow used for the adverse outcomes analytic part of the study. Schematic overview of the analytic pipeline. Each box represents a distinct step, and arrows represent dependencies between tasks. Preprocessing not depicted.

eFigure 2. Attrition diagram

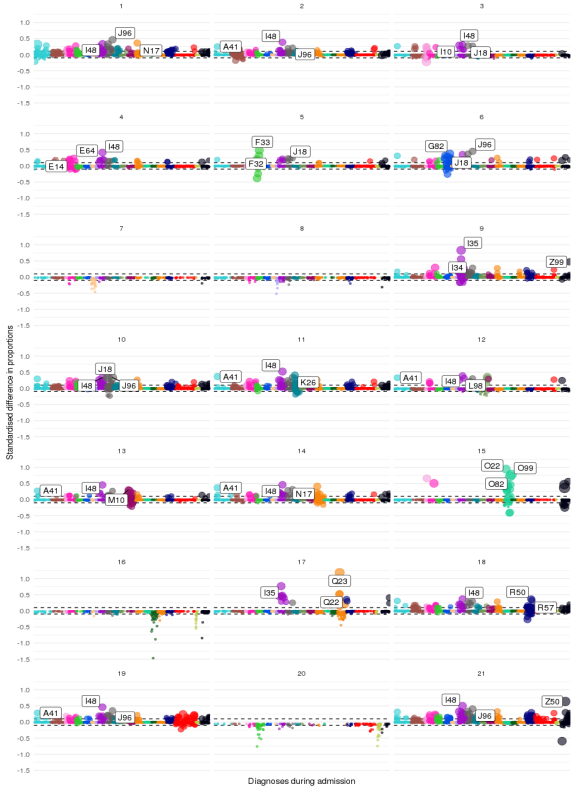


eFigure 3. Drugs prevalently involved in discouraged drug pairs



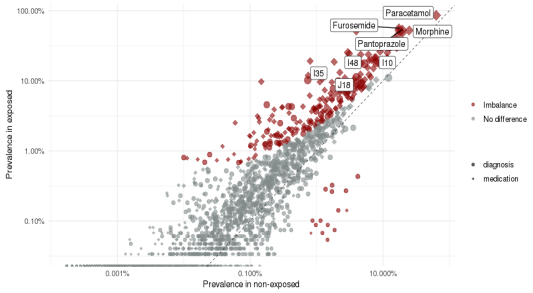
Drugs prevalently involved in discouraged drug pairs by management recommendation (columns) and clinical significance (rows), defined as drugs involved in discouraged drug pairs occurring in at least 10 of admissions. Points represent pDDIs, size the drug prevalence, and colour the clinical significance.

eFigure 4. Standardised differences in proportions of diagnoses in admissions with and without discouraged drug pairs by patient type



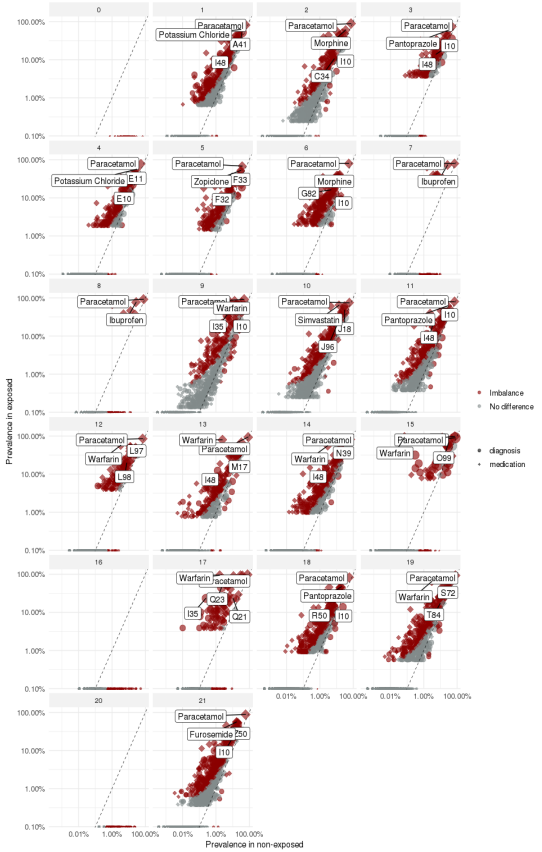
Standardised differences in proportions (i.e. discouraged drug pairs initiated versus not) of diagnoses and during admissions. The colour represents ICD-10 chapter, and the size is the prevalence in patients exposed to discouraged drug pairs. The top three diagnoses and drugs are labelled.

eFigure 5. Phenotyping of diagnoses and medications



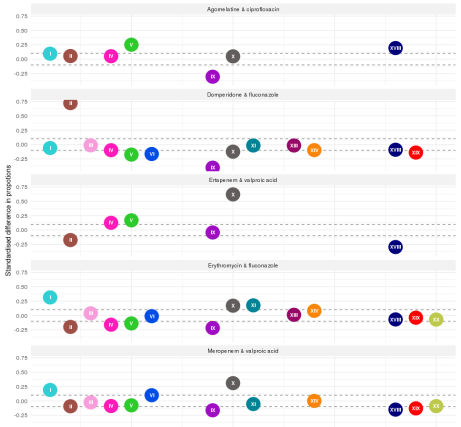
Detailed diagnoses and medication phenotyping of patients exposed to discouraged drug pairs vs. those not exposed. The plot compares the diagnoses and medications during hospitalisation. Each dot represents one of these covariates with the colour indicating the absolute value of the standardized difference in proportion (SPD). The top three medications and diagnoses with SPD >0.1 (exposed; red) are highlighted.

eFigure 6. Phenotyping of diagnoses and medications by patient type



Same as eFigure 5 but by patient type.

eFigure 7. Causes of death at ICD-10 chapter level among deceased exposed to selected drug pairs



Renal dysfunction and risk of inappropriate drug dosing

Full title

Identifying patients at high risk of inappropriate drug dosing in periods with renal dysfunction

Chapter contents

Identifying patients at high risk of inappropriate drug dosing in periods with renal dysfunction

Benjamin Skov Kaas-Hansen MD, MSc[*1,2] — Cristina Leal Rodríguez MSc[2]
Davide Placido MSc[2] — Hans-Christian Thorsen-Meyer MD[2,3]
Anna Pors Nielsen MD[2] — Nicolas Dérian MSc, PhD[4]
Søren Brunak MSc, PhD[2] — Stig Ejdrup Andersen MD, PhD[1]

Affiliations

[*] Corresponding author. Address: Munkesoevej 18, 4000 Roskilde, Denmark.
[1] Clinical Pharmacology Unit, Zealand University Hospital, Roskilde, Denmark
[2] NNF Center for Protein Research, University of Copenhagen, Denmark
[3] Department of Intensive Care Medicine, Copenhagen University Hospital (Rigshospitalet), Copenhagen, Denmark
[4] Data and Development Support, Region Zealand, Denmark

ORCID ID's

BSKH: 0000-0003-1023-0371. SEA: 0000-0002-1914-4720. CLR: 0000-0002-3133-0630. ND: 0000-0002-4477-023X. APN: 0000-0001-7903-5051. SB: 0000-0003-0316-5866.

Keywords

- Predictive modelling

- Kidney failure/Renal dysfunction
- Machine learning
- Risk markers
- Inappropriate drug dosing

Abstract

Introduction

Dosing of renally cleared drugs in patients with kidney failure often deviates from clinical guidelines but little is known about what is predictive of receiving inappropriate doses.

Methods and materials

We combined data from the Danish National Patient Register and in-hospital data on drug administrations and estimated glomerular filtration rates for admissions between 1 October 2009 and 1 June 2016, from a pool of about 2.9 million persons. We trained artificial neural network and linear logistic ridge regression models to predict the risk of five outcomes (>0, ≥1, ≥2, ≥3 and ≥5 inappropriate doses daily) with index set 24 hours after admission. We used time-series validation for evaluating discrimination, calibration, clinical utility and explanations.

Results

Of 52,451 admissions included, 42,250 (81%) were used for model development. The median age was 77 years; 50% of admissions were of women. ≥5 drugs were used between admission start and index in 23,124 admissions (44%); the most common drug classes were analgesics, systemic antibacterials, diuretics, antithrombotics, and antacids. The neural network models had better discriminative power (all AUROCs between 0.77 and 0.81) and were better calibrated than their linear counterparts. The main prediction drivers were use of anti-inflammatory, antidiabetic and anti-Parkison's drugs as well as having a diagnosis of chronic kidney failure. Sex and age affected predictions but slightly.

Conclusion

Our models can flag patients at high risk of receiving at least one inappropriate dose daily in a controlled in-silico setting. A prospective clinical study may confirm this holds in real-life settings and translates into benefits in hard endpoints.

Introduction

Renal diseases affect patients' susceptibility to, and modify the effects of many drugs, and they reduce renal clearance exposing patients to higher steady-state concentrations when given standard doses. The kidneys excrete active forms and/or metabolites of many drugs, so renal dysfunction necessitates dose-adjustment of renally cleared drugs with narrow therapeutic indices to prevent adverse events and accidental over-dosing.

Inadequate dose-adjustment of such drugs has been linked to polypharmacy [1,2] and can cause noxious events [3] or accidental over-dosing [4]. Although not a new issue, [5,6] deviating from guidelines is widespread with prevalence estimates up to 70% [1,2,7-9]. Despite large inter-individual variability in clearance and response, dose adjustment for many drugs is crude and based on the estimated glomerular filtration rate (eGFR), for example, halving the dose when eGFR < 60 ml/min/1.73 m^2.

Appropriate alerts in order-entry systems may facilitate rational clinical decision-making, [10,11] and convincing examples have showcased how computerised systems can underpin rational pharmacotherapy [4,12]. However, downsides of extensive computerisation of healthcare emerge [13]; alert fatigue [14] is particularly problematic, and strategies and interventions have been proposed to mitigate its negative effects [15].

At Danish hospitals, prescriptions are mostly dispensed and administered by nurses who record detailed meta-data [16]. Prescriptions are usually made and revised by physicians regularly during clinical rounds, typically in the morning or early afternoon. Electronic decision support is generally immature and neither prescribing physicians nor dispensing nurses are warned if dose-adjustment be advised or even required.

We suspect that the need for dose-adjustment in patients with renal dysfunction often goes unrecognised. Thus, with this paper we study its predictability to inform clinicians and healthcare personnel upfront about which patients with renal dysfunction are at elevated risk of inappropriate drug dosing. To this end we used and compared predictive modelling methods from classical statistical modelling and machine learning.

Methods

Study design, patients and data

We conducted a register-based prediction study with prospective data for patients admitted to 12 public hospitals in two Danish regions comprising about 2.9 million persons (more than half the Danish population). We collected diagnosis data from the Danish National Patient Register, demographic data from the Danish Civil Registration System [17], as well as medication and biochemical data from electronic patient records. Diagnoses were encoded using the 10th revision of the International Classification of Diseases (ICD-10), drugs with the Anatomical and Therapeutic Chemical classification (ATC).

The units of analysis were inpatient admissions, defined as chains of successive in-hospital visits at most 24 hours apart. We included admissions starting between 1 October 2009 and 1 June 2016, with at least one eGFR measurement ≤30 during the first 24 hours of admission. We excluded minors (age <18 years). Admission time uses hour resolution (an admission starting at 9:54 is recorded as starting at 9:00) so to ensure at least 24 hours of observation time before inclusion, index was set at hour of admission + 25 hours. Prior sample-size estimation was foregone.

Outcomes

The outcome variables were based on the daily rate = r/E of inappropriate doses during follow-up, capped at 30 days. r is the number of given inappropriate doses of select drugs cleared mainly renally and with narrow therapeutic indices; E the time-at-risk (figure 1). To obtain well-defined times-at-risk, we set the eGFR threshold to ≤30 ml/min/1.73m^2 (unit omitted from here onward) and used the rules in supplementary table S1 for counting the number of inappropriate doses, based on the official reference guidelines for Danish physicians (pro.medicin.dk) as of January 2021.

We used two rules, one definitive (maximum daily dose = 0 mg) and one of dose-adjustment (reduced daily dose). Operationalisation of the definitive rule is straightforward: if the last eGFR ≤30, there should be no administrations until an eGFR >30 is measured. The dose-adjustment rule is slightly more involved as inappropriate dosing comes in two forms: (a) on a given day there are more than one eGFR measurements, of which at least one is ≤30, and the cumulative daily dose surpasses the threshold in the period(s) between above-threshold measurements, or (b) all eGFR measurements of a given day are ≤30 and the cumulative daily dose surpasses the threshold.

Variables and features

Variables are original data (e.g. sex and age at admission) and features the results of rendering the variables appropriate as model inputs (e.g. one-hot-encoded day of admission). Based on clinical and pharmacological experience we hand-picked pertinent variables likely to be informative to the prediction problem and realistically available in the clinical setting. These fall into three categories. Demographic: age at admission

(numeric), sex (binary). Clinical: number of distinct drugs (ATC level 5) administered between admission and index (numeric); therapeutic drug classes (ATC level 2) used between admission and index (one-hot-encoded); the Elixhauser score at admission (numeric, AQHR adaptation) [18]; ICD-10 chapters of diagnoses recorded in the past five years before admission (one-hot-encoded); record of chronic kidney failure in the past five years before admission (ICD-10 N18* diagnoses, one-hot-encoded). Contextual: hour of admission (numeric, transformed as *f(t) = abs(12 – t);* see supplementary figure S1); weekday of admission (one-hot-encoded); number of admissions in the past 5 years before admission (numeric).

Missing values, only present for hour of admission and discharge, were imputed by sampling from the empirical distributions of valid values.

Models and training

We tried two model architectures (linear logistic ridge regression and artificial neural network) with several binary outcomes defined by increasing thresholds of the daily rate of inappropriate doses (>0, ≥1, ≥2, ≥3 and ≥5). The neural network models were multilayer perceptrons (MLPs) enabling speedy training and evaluation.

All admissions starting before 1 July 2015 were assigned to the development set (42,250 admissions [81%] of 27,253 patients) and the rest to the independent hold-out test set (10,201 admissions [19%] of 8,412 patients). Because admissions constitute the unit of analysis, some patients likely appear in both the development and test sets. Information may leak between the sets [19] so as a sensitivity analysis, we evaluated the performance also in the subset of test-set patients not in the development set.

We used the multivariate *TPEsampler* from O*ptuna* [20] to find the best-performing hyperparameters by sampling 100 configurations, each using 5-fold stratified-and-grouped cross-validation, from the following proposal distributions (discrete values in round brackets, bounds of log-uniform distributions in squared): optimiser (Adam, RMSprop), learning rate [10^{-6}, 10^{-1}], activation function (tanh, sigmoid), L2 penalty [10^{-6}, 10^{-2}], number of hidden layers (1, 2, 3, 4), number of nodes per hidden layer [16, 32, 65, 128], batch size (32, 64, 128, 256, 512), class handling (see below).

Only relevant hyperparameters were sampled and we ran Optuna on linear and MLP models separately because they have disparate hyperparameter sets. MLP models with more hidden layers and more nodes therein can learn more complex relationships but become prone to overfitting which we countered with early stopping [21] and L2 regularisation (handles collinearity better than L1 regularisation) [22,23]. The batch size is the number of observations from which the model learns at a time; small batches can give outliers undue influence while full-batch training (batch size = number of units) can become computationally impractical [19]. Class imbalances in binary outcomes can misguide training, so we tested the following remedies: synthetic minority oversampling technique (SMOTE), random over-sampling of minority class, NearMiss, random under-sampling of majority class, class weighting, and none. SMOTE creates a dataset similar to the minority class but of the same size as the majority class [24]; NearMiss downsizes the majority class in a systematic way to retain as much information as possible in fewer data points [25]. Class weighting retains the original data but gives more weight to minority-class observations.

Hyperparameter optimisation models trained for maximum 500 epochs with 50-epoch patience on improvement in the validation loss. The final models were trained on the full development set until the loss reached that obtained in the best cross-validation fold for the best configuration [21].

Evaluation and explanation

Discrimination was assessed with receiver operating characteristic (ROC) curves and areas under the ROC curves (AUROC), calibration-in-the-small by plotting decile-binned predicted probabilities against corresponding bin-wise observed event proportions [26] with 95% Jeffrey intervals [27]; results from a perfectly calibrated model fall on the diagonal. We used the decision-curve analytic framework to gauge the models' potential clinical utility [28,29].

For explanation and scrutiny of prediction drivers, we used the SHAP DeepExplainer yielding one shap value per feature per unit [30]. The shap value for a risk prediction model is the absolute change in risk of a given unit's value for each feature: the cohort-wide mean risk plus the sum of one unit's shap values equals that unit's risk.

Analysis and ethics

The full analytical pipeline was built with Snakemake [31] (schematic overview in supplementary figure S2) to facilitate transparency and reproducibility; blinding was impractical and so foregone, but all analytic code is available online (DOI: 10.5281/zenodo.4560078). Univariate distributions were summarised by median (inter-quartile range) and count (proportion), as appropriate. This report adheres to pertinent items in the MINIMAR guideline [32] and TRIPOD statement [33].

All data have been marshalled on Computerome, a secure high-performance Danish computing infrastructure, after obtaining approval from the Danish Patient Safety Authority (3-3013-1723; then competent authority for ethical approval), the Danish Data Protection Agency (DT SUND 2016-48, 2016-50, 2017-57) and the Danish Health Data Authority (FSEID 00003724). Results

Table 1 shows univariate summary statistics of the 52,451 admissions (42,250 + 10,201) of 35,665 patients (27,253 + 8,412) included in the study (see supplementary table S2 for extended version with all features). Patients in the test sets were similar to those in the development set with some notable exceptions. Fewer had received inappropriate doses, especially in the test-set patients not part of the development set who also had fewer previous admissions.

In the development set, the median age was 77 years (IQR: 67-85) and 20,743 admissions (49%) were of 13,759 women (50%). The median time at risk was 3.5 days (inter-quartile range: 1.7–7.7) and at least one inappropriate dose was given in 3,786 admissions (9.0%); ≥1 inappropriate dose daily was given in 5.3% of admissions and ≥5 inappropriate doses daily were given in 0.9%. The target drugs most commonly given in inappropriate doses were ibuprofen (M01AE01, 4.1%) and metformin (A10BA02, 3.4%); inappropriate doses of the other target drugs were given in <1% of admissions.

Patients in 4,988 admissions (12%) had no admissions in the 5 years before inclusion; 13,960 (33%) had ≥7 previous admissions. The most common drug classes used between admission and index were analgesics (N02, 37%), systemic antibacterials (J01, 35%), diuretics (C03, 33%) antithrombotics (B01, 28%), and antacids (A02, 25%). Previous diagnoses were most commonly cardiovascular (chapter IX, 61%), genitourinary (XIV,

55%), related to i.a. lesions and external causes (XIX, 48%), endocrine-metabolic (IV, 47%), and symptoms/abnormal findings (XVIII, 44%).

Table 2 shows the hyperparameters of the best configurations with performance metrics of the final models (see also supplementary figures S3–S12). Generally, multi-layer perceptron (MLP) models performed slightly better than their linear counterparts, all obtaining AUROC's between 0.77 and 0.81 in the test set (ROC curves in supplementary figures S13–S22). The MLP models more consistently showed good calibration in the development set. For daily rates >0, ≥1 and ≥2 both MLP and linear models were very well-calibrated in the test set (supplementary figures S23–S32). The decision curves did not suggest the clinical utility of the MLP models be superior to that of the linear (supplementary figures S33–S42).

The model-specific shap values offer some insights (supplementary figures S43–S53). First, many features contribute substantively to the predictions of daily rate >0 and ≥1 outcomes, while few features almost entirely drive the predictions for the other outcomes. Second, few features are the dominant prediction drivers across outcomes and models: use of anti-inflammatory, antirheumatic and antidiabetic drugs as well as diagnoses of chronic kidney failure. Third, sex and age contribute little to predictions. Fourth, using more distinct drugs (reflecting various levels of polypharmacy) pushes the risk up and using fewer drugs pulls the risk down. Fifth, the linear models tend to give most weight to relatively few features whereas the MLP models spread out the contributions across more features. Finally, the number of previous admissions (a proxy for frailty) became an increasingly important driver with increasing rarity of the outcome, in the MLP models.

Figure 2 shows the relationships between values of select features and their shap values and illustrates how MLP models capture highly non-linear effects and near-linear effects as appropriate (e.g. the effects of age at admission and number of previous admissions for daily rate >0.)

Discussion

This study reveals that 9.0% of patients with reduced kidney function are exposed to inappropriate doses of selected renal risk drugs in the follow-up period. Our models performed quite well with AUROC's between 0.77 and 0.81 with good calibration-in-the-small for daily rates >0 and ≥1, in the test set. For rarer outcomes (daily rates ≥2, ≥3 and ≥5) calibration suffered and clinical utility is unlikely to be substantive.

Apt intervention necessitates comprehension of the nature and extent of the problem. Use of renal risk drugs and associated problems, including inappropriate dosing, in patients with renal dysfunction is well-described [34-38]. A cross-sectional study of 83,000 American outpatient Veterans found that 32% of patients with creatinine clearance between 15 and 29 were given drugs at excessive doses considering their kidney function [39]. Medication burden had the strongest cooccurrence with inappropriate dosing and metformin was a prominent drug among those with inappropriate doses. This agrees with our findings although our study design has clearer temporality.

Some have called for a prediction tool to identify elderly at elevated risk of adverse drug reactions [40], a notion similar to ours in spirit but different in scope. Studies of factors associated with inadequate dose adjustment are few and often of retrospective nature eliciting relationships with characteristics after inappropriate doses have already been given. One study seeking to elicit factors associated with dosing appropriateness, using a

logistic regression, reported the statistically strongest association to be with severity of chronic kidney failure (p-value = 7%) [41]. A similar study found dosing errors in 33% of the patients; *age* (odds ratio, OR: 1.05), *number of drug prescriptions* (OR: 1.1) and *number of drugs requiring dose adjustment* (OR: 2.0) were associated with dosing errors [42]. A third study found that, in patients with chronic kidney failure, *late-stage chronic kidney disease*, *number of prescribed drugs* and *presence of comorbidity* were associated with dosing errors. Ill-defined indices and times-at-risk render such enquiries of little use for a priori prediction and risk stratification: the ability to intervene presupposes a reliable estimate of risk in advance, before the event happens.

Carey et al. found only few factors to be genuinely predictive of potentially inappropriate prescribing in elderly outside the hospital setting [43]. Our models had AUROC's (0.77–0.81) slightly higher than that of their model (0.76). In a prospective study from Norway [35] of internal-medicine patients with a mean age of 71 years, 35% received suboptimal doses; a composite variable (*number of clinical/pharmacological risk factors*) was quite strongly associated with non-optimal dosing (RR: 1.33), less so *number of drugs at admission* (RR: 1.09), whereas *sex* and *age* were not predictive of non-optimal dosing. Our results agree quite well with that finding, probably because the information captured by age and sex (essentially, proxies of comorbidity) is expressed explicitly in our feature set.

As such, our models fare quite well with performance metrics superior to those of other published models even though ours came from an independent and temporally distinct test set. Many studies employing machine learning models for predicting medical outcomes use normal split-sample validation, putting aside a random sample of the observations for testing. This has several logical and practical implications, perhaps most

notably that a model developed with data collected between, say, 2005 and 2015 will likely perform better in a test case from 2013 than in one from 2017. The subset of our test set with patients not part of the development set is a conceptually appealing way to gauge how the model might perform in a new population. It does, however, distort the data and somewhat delink it from the clinical reality: some patients have previous admissions and those admitted for the first time are probably different from the rest.

Strengths

Here we highlight five principal strengths of this study. First, this is by far the largest study of its kind to date. Second, time-series validation yielded realistic performance evaluation in distinct (future) data [44] vis-a-vis many articles on predictive modelling, perhaps most clearly seen in the surge of COVID-19 papers [45]. Third, our data were richer than in any other study in this area thanks to the combined diversity and reliability of longitudinal diagnostic data from the National Patient Register and deep phenotypic in-hospital data. Fourth, our summary statistics are well-aligned with descriptive studies of deviations from dosing recommendations, and the nature of the general patient population to which a model as ours would be applied [46]. Finally, the shap-value analysis suggests that the models picked up clinically relevant information without undue influence of individual predictors.

Limitations

Like any study, this has potential limitations. First, albeit simple and elegant, using *only* eGFR as a proxy for kidney function is not always advisable [47]. It is, however, considered a reasonable metric for medicinal dosing [48] and used in Danish guidelines.

Second, eGFR can be estimated in several ways [49] and both the 4-variable MRDR Study and CKD-EPI equations were used in our data. However, clinicians use the reported eGFR estimate as-is and both equations perform well for low eGFR values [50]. Third, hard thresholds on eGFR are arbitrary: the difference in kidney function between eGFRs of 29 and 31 is minuscule, but the cutoff must be set somewhere. Again, we stayed loyal to the guidelines as these are, nevertheless, what should support clinicians' prescribing decisions. Fourth, many drugs have narrow and intermediate therapeutic indices. We focused on seven drugs cleared primarily by the kidneys and with narrow therapeutic indices that are fairly common in a Danish setting and span several important drug classes. The drugs included also allowed for reasonably harmonised rules of inappropriate dosing. Finally, our binary outcomes are soft endpoints and do constitute a simplification. Seemingly inappropriate doses could be conscious choices and the outcome variables do not capture information about actual toxicity experienced by the patient. However, the narrow therapeutic indices of the included drugs increase the likelihood of noxious effects without appropriate dose adjustment.

Conclusion

Despite physicians' awareness of the need for dose adjustment in patients with kidney dysfunction, a well-performing clinical decision support tool may help prevent such patients from "flying under the radar" in a busy clinical setting. Indeed, our models can flag patients at high risk of receiving >0 or ≥1 inappropriate dose daily.

A prospective evaluation is necessary to assess if these results transport to the clinic and if the models can offer genuine clinical utility for the patients. Receiving inappropriate doses is a soft endpoint, so clinical evaluation should consider also hard endpoints, either

generic (e.g. length-of-stay, need for post-discharge rehabilitation and mortality) or specific ones related to the target drugs (e.g. transfusion and occurrence of known side-effects of these drugs).

Data availability

Due to the sensitive nature of the data, we can neither offer access to nor share our data with third parties. Data can be obtained from the original sources upon request.

Acknowledgements

The authors would like thank Innovation Fund Denmark (5153-00002B) and the Novo Nordisk Foundation (NNF14CC0001, NNF17OC0027594) for their financial contribution to BigTempHealth without which this study had not been possible. The funders played no role in designing, conducting, interpreting, or reporting this study.

Contributions

Conceptualisation: BSKH, SEA. Data curation: BSKH, CLR. Formal analysis: BSKH. Methodology: APN, BSKH, DP, HCTM, ND. Software: BSKH. Code review: CLR, DP, HCTM. Drafting: BSKH. Funding acquisition: SB, SEA. Resources: SB, SEA. Supervision: SEA. Review: All.

Conflicts of interest

The authors declare the following competing interests:

- BSKH: None
- CRL: None

- DP: None
- HCTM: None
- ND: None
- APN: None
- SB reports ownerships in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S, and managing board memberships in Proscion A/S and Intomics A/S outside the submitted work
- SEA: None

Tables

Table 1: Univariate summary statistics of select features. Values are median (inter-quartile range) and count (proportion) as appropriate. *Distinct patients* and *Distinct women* show counts of actual patients (as a patient can contribute more than one unit.)

Variate	Development set (N = 42,250)	Test set (N = 10,201)	Test set (not in devel. set) (N = 5,980)
Women	20,743 (49%)	4,854 (48%)	2,940 (49%)
Distinct patients	27,253	8,412	5,341
Distinct women	13,759 (50%)	4,049 (48%)	2,629 (49%)
Time at risk, days	3.5 (1.7–7.7)	3.5 (1.7–7.2)	2.9 (1.5–6.4)
Inappropriate doses (outcomes)			
> 0 (at least one)	3,786 (9.0%)	1,080 (11%)	740 (12%)
≥ 1 daily	2,241 (5.3%)	588 (5.8%)	333 (5.6%)
≥ 2 daily	1,236 (2.9%)	288 (2.8%)	108 (1.8%)
≥ 3 daily	783 (1.9%)	171 (1.7%)	56 (0.9%)
≥ 5 daily	366 (0.9%)	64 (0.6%)	9 (0.2%)
Admissions 5 years before admission			
None	4,988 (12%)	1,082 (11%)	1,074 (18%)
1–2	10,100 (24%)	2,367 (23%)	1,873 (31%)
3–4	7,712 (18%)	1,919 (19%)	1,232 (21%)
5–6	5,490 (13%)	1,303 (13%)	685 (12%)
≥ 7	13,960 (33%)	3,530 (35%)	1,116 (19%)
Drugs used between admission and index			
None	6,165 (15%)	1,228 (12%)	762 (13%)
1–2	9,111 (22%)	1,984 (19%)	1,254 (21%)
3–4	8,761 (21%)	2,078 (20%)	1,355 (23%)
5–6	7,197 (17%)	1,852 (18%)	1,095 (18%)
≥ 7	11,016 (26%)	3,059 (30%)	1,514 (25%)
Any diagnosis of chronic kidney failure	13,470 (32%)	3,391 (33%)	732 (12%)
Top-5 ICD-10 chapters†			
Cardiovascular (IX)	25,757 (61%)	6,392 (63%)	3,283 (55%)
Genitourinary (XIV)	23,025 (55%)	5,819 (57%)	2,306 (39%)
Lesions, external causes, etc. (XIX)	20,275 (48%)	4,749 (47%)	2,481 (42%)
Metabolic-endocrine (IV)	19,716 (47%)	5,096 (50%)	2,415 (40%)
Symptoms/abnormal findings (XVIII)	18,663 (44%)	5,711 (56%)	2,882 (48%)
Top-5 drug classes‡			

Analgesics (N02)	15,740 (37%)	4,367 (43%)	2,506 (42%)
Systemic antibacterials (J01)	14,719 (35%)	3,257 (32%)	1,938 (32%)
Diuretics (C03)	13,966 (33%)	3,672 (36%)	1,951 (33%)
Antithrombotics (B01)	11,842 (28%)	3,181 (31%)	1,795 (30%)
Antacids (A02)	10,635 (25%)	2,776 (27%)	1,407 (24%)

† ICD-10 chapters (Roman numbering) of diagnoses recorded in the last 5 years before admission.
‡ Drug classes (ATC level 2) administered between admission and index.

Table 2: Performance metrics of final models and results of Optuna hyperparameter optimisation. AUROC: area under the receiver operating characteristic curve. MLP: multi-layer perceptron. Undersample: random sample of the size of the minority class, from the majority class. Oversample: randomly sample (with replacement) from the minority class until reaching a sample size equal to the size of the majority class. SMOTE: synthetic minority oversampling technique [24]. NearMiss: a method for non-random, systematic downsampling of the majority class while retaining as much information as possible [25].

Parameter	Daily rate >0		Daily rate ≥1		Daily rate ≥2		Daily rate ≥3		Daily rate ≥5	
	Linear	MLP	Linear	MLP	Linear	MLP	Linear	MLP	Linear	MLP
AUROC										
Development set	0.80	0.81	0.81	0.83	0.81	0.84	0.82	0.83	0.82	0.83
Test set	0.77	0.79	0.78	0.79	0.79	0.79	0.81	0.81	0.78	0.80
Test set (new patients)	0.78	0.79	0.82	0.83	0.86	0.86	0.89	0.90	0.82	0.79
Hyperparameters										
Batch size	512	128	512	32	32	64	256	256	64	64
Class handling	Undersample	SMOTE	NearMiss	NearMiss	Oversample	SMOTE	Oversample	NearMiss	Oversample	None
L2 penalty	1.28×10^{-6}	1.66×10^{-6}	3.02×10^{-6}	1.43×10^{-6}	4.38×10^{-6}	1.39×10^{-6}	1.43×10^{-6}	1.30×10^{-6}	1.09×10^{-5}	3.94×10^{-6}
Learning rate	1.79×10^{-2}	1.20×10^{-4}	1.92×10^{-2}	3.45×10^{-4}	6.73×10^{-3}	2.71×10^{-4}	3.76×10^{-2}	3.08×10^{-4}	2.11×10^{-2}	4.86×10^{-4}
Optimiser	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Activation function	—	tanh	—	sigmoid	—	tanh	—	sigmoid	—	sigmoid
No. hidden layers	—	3	—	1	—	1	—	1	—	2
Nodes per hidden layer	—	8	—	8	—	32	—	32	—	8

352

Figures

Figure 1: Deriving the outcome variables. This exemplary admission is composed of three successive in-patient visits (i.e. the patient has been transferred twice represented by the arrows). The admission is eligible because it spans more than 24 hours and an eGFR ≤30 was measured before index. Here, apixaban was given while the patient's eGFR was ≤30, but dose reduction rendered these administrations appropriate.

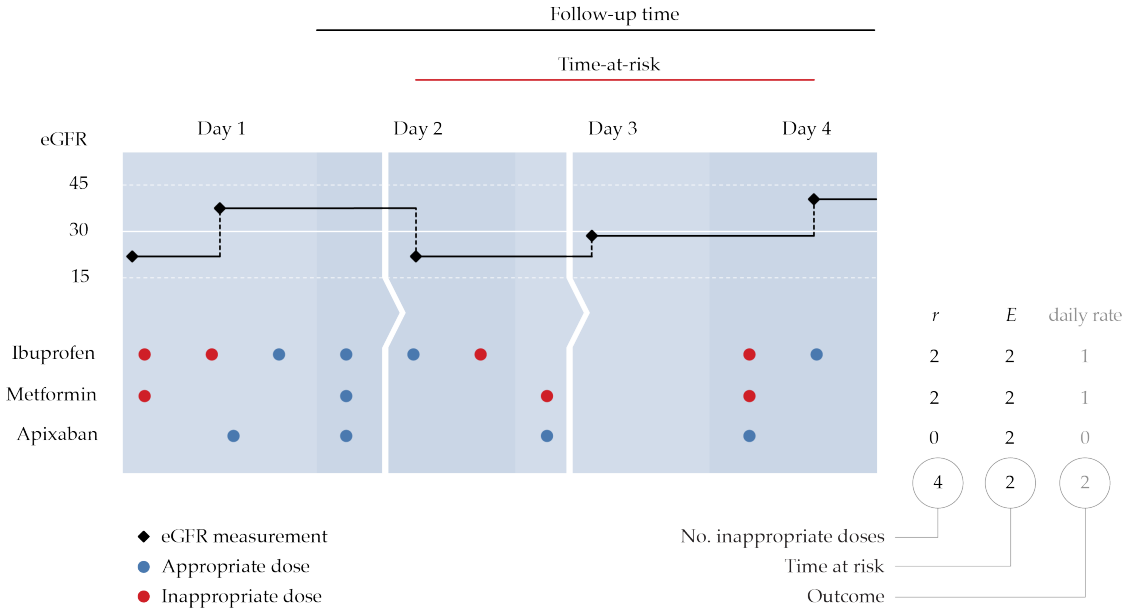
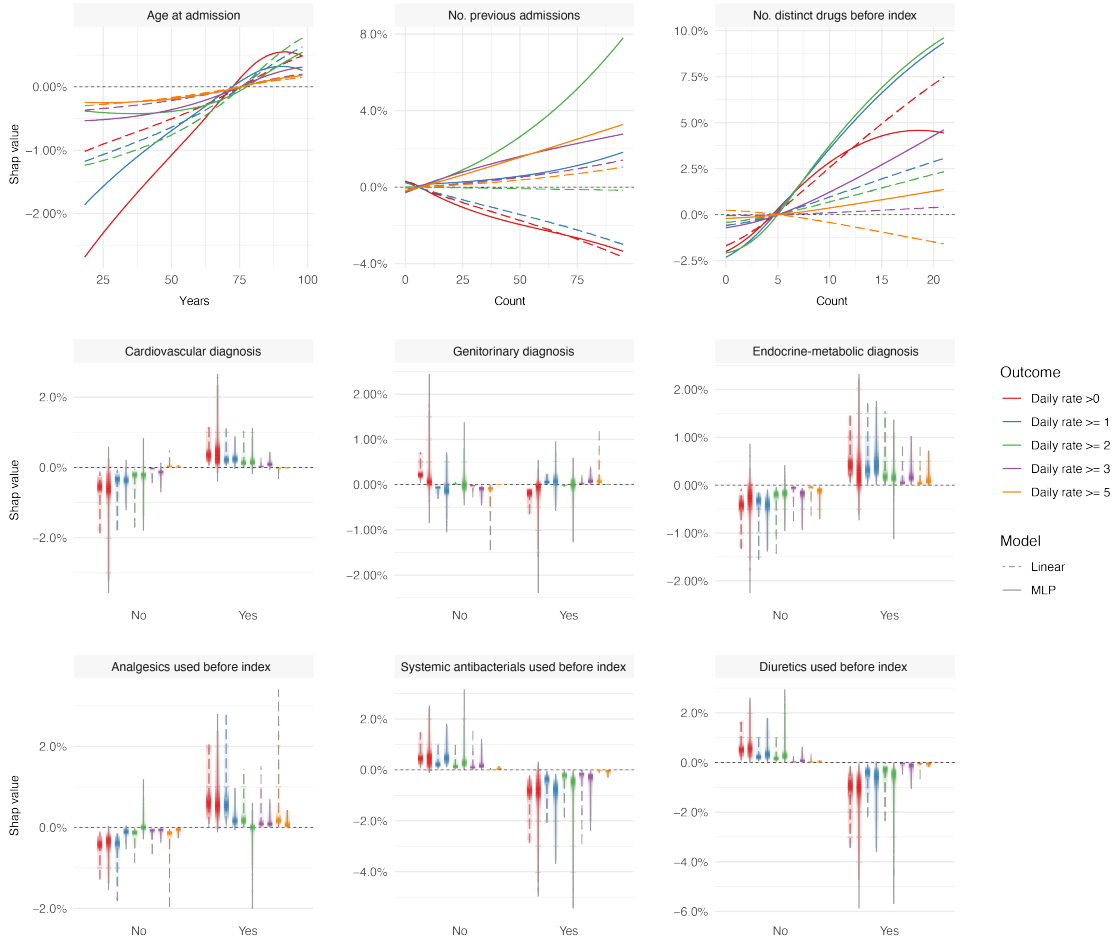


Figure 2: Bivariate relationships between values of select features (x axis) and their corresponding shap values (y axis). The continuous features are summarised by locally estimated scatterplot smoothing (LOESS), binary features by vertical density bands.



References

1. Saleem, A. & Masood, I. Pattern and Predictors of Medication Dosing Errors in Chronic Kidney Disease Patients in Pakistan: A Single Center Retrospective Analysis. *PLoS One* **11,** e0158677 (2016).
2. Hoffmann, F. et al. Renal Insufficiency and Medication in Nursing Home Residents. A Cross-Sectional Study (IMREN). *Dtsch. Arztebl. Int.* **113,** 92–98 (2016).
3. Munar, M. Y. & Singh, H. Drug dosing adjustments in patients with chronic kidney disease. *Am. Fam. Physician* **75,** 1487–1496 (2007).
4. Niedrig, D. et al. Development, implementation and outcome analysis of semi-automated alerts for metformin dose adjustment in hospitalized patients with renal impairment. *Pharmacoepidemiol. Drug. Saf.* **25,** 1204–1209 (2016).
5. Bernstein, J. M. & Erk, S. D. Choice of antibiotics, pharmacokinetics, and dose adjustments in acute and chronic renal failure. *Med. Clin. North. Am.* **74,** 1059–1076 (1990).
6. Khare, A. K. Antibiotic dose adjustment in renal insufficiency. *Lancet* **340,** 1480 (1992).
7. Dorks, M., Allers, K., Schmiemann, G., Herget-Rosenthal, S. & Hoffmann, F. Inappropriate Medication in Non-Hospitalized Patients With Renal Insufficiency: A Systematic Review. *J. Am. Geriatr. Soc.* **65,** 853–862 (2017).
8. Getachew, H., Tadesse, Y. & Shibeshi, W. Drug dosage adjustment in hospitalized patients with renal impairment at Tikur Anbessa specialized hospital, Addis Ababa, Ethiopia. *BMC Nephrol.* **16,** 158 (2015).
9. Altunbas, G. et al. Renal Drug Dosage Adjustment According to Estimated Creatinine Clearance in Hospitalized Patients With Heart Failure. *Am. J. Ther.* **23,** e1004-8 (2016).
10. Hillestad, R. et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff. (Millwood)* **24,** 1103–1117 (2005).
11. Stewart, W. F., Shah, N. R., Selna, M. J., Paulus, R. A. & Walker, J. M. Bridging the inferential gap: the electronic health record and clinical evidence. *Health Aff. (Millwood)* **26,** w181-91 (2007).
12. Boussadi, A. et al. Validity of a clinical decision rule-based alert system for drug dose adjustment in patients with renal failure intended to improve pharmacists' analysis of medication orders in hospitals. *Int. J. Med. Inform.* **82,** 964–972 (2013).
13. Gawande, A. Why doctors hate their computers. *The New Yorker* (2018).
14. Baysari, M. T., Tariq, A., Day, R. O. & Westbrook, J. I. Alert override as a habitual behavior - a new perspective on a persistent problem. *J. Am. Med. Inform. Assoc.* **24,** 409–412 (2017).
15. Kane-Gill, S. L. et al. Technologic Distractions (Part 1): Summary of Approaches to Manage Alert Quantity With Intent to Reduce Alert Fatigue and Suggestions for Alert Fatigue Metrics. *Crit. Care Med.* **45,** 1481–1488 (2017).
16. Jensen, T. B. et al. Content and validation of the Electronic Patient Medication module (EPM)—the administrative in-hospital drug use database in the Capital Region of Denmark. *Scand. J. Public Health* **0,** 1403494818760050 (2018).
17. Schmidt, M. et al. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin. Epidemiol.* **7,** 449–490 (2015).
18. Moore, B. J., White, S., Washington, R., Coenen, N. & Elixhauser, A. Identifying Increased Risk of Readmission and In-hospital Mortality Using Hospital Administrative Data: The AHRQ Elixhauser Comorbidity Index. *Med. Care* **55,** 698–705 (2017).
19. Chollet, F. *Deep Learning with Python* (Manning Publications Co., New York, USA, 2018).
20. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Preprint at http://arxiv.org/abs/1907.10902 (2019).
21. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge (MA), USA, 2016).
22. Efron, B. & Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Cambridge University Press, London, United Kingdom, 2016).
23. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed., Springer, New York, 2009).
24. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16,** 321–357 (2002).

25. Zhang, J. & Mani, I. kNN approach to unbalanced data distributions: a case study involving information extraction. In: *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets* (2003).
26. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating* (Springer, New York, 2009).
27. Brown, L. D., Cai, T. T. & DasGupta, A. Interval Estimation for a Binomial Proportion. *Statist. Sci.* **16,** 101–133 (2001).
28. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making.* **26,** 565–574 (2006).
29. Kerr, K. F., Brown, M. D., Zhu, K. & Janes, H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J. Clin. Oncol.* **34,** 2534–2540 (2016).
30. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems 30* (2017).
31. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).
32. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A. & Shah, N. H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Inform. Assoc.* (2020).
33. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* **162,** 55–63 (2015).
34. Saad, R., Hallit, S. & Chahine, B. Evaluation of renal drug dosing adjustment in chronic kidney disease patients at two university hospitals in Lebanon. *Pharm. Pract. (Granada)* **17,** (2019).
35. Blix, H. S. et al. The majority of hospitalised patients have drug-related problems: results from a prospective study in general hospitals. *Eur. J. Clin. Pharmacol.* **60,** 651–658 (2004).
36. Andreu Cayuelas, J. M. et al. Kidney function monitoring and nonvitamin K oral anticoagulant dosage in atrial fibrillation. *Eur. J. Clin. Invest.* **48,** e12907 (2018).
37. Seiberth, S. et al. Correct use of non-indexed eGFR for drug dosing and renal drug-related problems at hospital admission. *Eur. J. Clin. Pharmacol.* (2020).
38. Breton, G. et al. Inappropriate drug use and mortality in community-dwelling elderly with impaired kidney function–the Three-City population-based study. *Nephrol. Dial. Transplant.* **26,** 2852–2859 (2011).
39. Chang, F., O'Hare, A. M., Miao, Y. & Steinman, M. A. Use of Renally Inappropriate Medications in Older Veterans: A National Study. *J. Am. Geriatr. Soc.* **63,** 2290–2297 (2015).
40. Parameswaran Nair, N. et al. Hospitalization in older patients due to adverse drug reactions - the need for a prediction tool. *Clin. Interv. Aging* **11,** 497–505 (2016).
41. Kalender-Rich, J. L., Mahnken, J. D., Wetmore, J. B. & Rigler, S. K. Transient impact of automated glomerular filtration rate reporting on drug dosing for hospitalized older adults with concealed renal insufficiency. *Am. J. Geriatr. Pharmacother.* **9,** 320–327 (2011).
42. Won, H.-J. et al. Evaluation of medication dosing errors in elderly patients with renal impairment. *Int. J. Clin. Pharmacol. Ther.* **56,** 358–365 (2018).
43. Carey, I. M. et al. What Factors Predict Potentially Inappropriate Primary Care Prescribing in Older People? *Drug Aging* **25,** 693–706 (2008).
44. Steyerberg, E. W. & Harrell, F. E. J. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* **69,** 245–247 (2016).
45. Wynants, L. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369,** (2020).
46. Yusuf, M. et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open* **10,** (2020).
47. Eppenga, W. L. et al. Drug therapy management in patients with renal impairment: how to use creatinine-based formulas in clinical practice. *Eur. J. Clin. Pharmacol.* **72,** 1433–1439 (2016).
48. Rule, A. D. & Glassock, R. J. GFR estimating equations: getting closer to the truth? *Clin. J. Am. Soc. Nephrol.* **8,** 1414–1420 (2013).
49. Corsonello, A. et al. Estimating renal function to reduce the risk of adverse drug reactions. *Drug Saf.* **35 Suppl 1,** 47–54 (2012).
50. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150,** 604–612 (2009).

Identifying patients at high risk of inappropriate drug dosing in periods with renal dysfunction

Supplement

Benjamin Skov Kaas-Hansen

19 Feburary 2021

Contents

This document contains all supplementary tables and figures for the above-mentioned paper. The figures within each domain all share the same caption, and so to avoid cluttering captions are given only once, under the domain heading.

Supplementary tables

Table S1: Maximum doses in periods when eGFR is as indicated. ATC codes in brackets.

	Drug	eGFR thredshold	Max. daily dose
1	Apixaban (B01AF02)	<= 30	5 mg.
2	Dabigatran (B01AE07)	<= 30	0 mg.
3	Rivaroxaban (B01AF01)	<= 30	0 mg.
4	Metformin (A10BA02)	<= 30	0 mg.
5	Ibuprofene (M01AE01)	<= 30	0 mg.
6	Celecoxib (M01AH01)	<= 30	0 mg.

7	Methotrexate (L04AX03)	<= 30	0 mg.
8	Lithium citrate (N05AN01)	<= 30	0 mg.

Table S2 (extended version of table 1): Univariate summary statistics of the three data sets. Values are median (inter-quartile range) and N (%).

	variate	develop	test	test_new	in_table1
1	n	42250	10201	5980	yes
2	time_at_risk (median [IQR])	3.5 [1.7, 7.7]	3.5 [1.7, 7.2]	2.9 [1.5, 6.4]	yes
3	daily_rate_not_zero (%)	3786 (9.0)	1080 (10.6)	740 (12.4)	yes
4	daily_rate_geq_1 (%)	2241 (5.3)	588 (5.8)	333 (5.6)	yes
5	daily_rate_geq_2 (%)	1236 (2.9)	288 (2.8)	108 (1.8)	yes
6	daily_rate_geq_3 (%)	783 (1.9)	171 (1.7)	56 (0.9)	yes
7	daily_rate_geq_5 (%)	366 (0.9)	64 (0.6)	9 (0.2)	yes
8	inapp_dose_of_M01AE01 (%)	1713 (4.1)	343 (3.4)	244 (4.1)	yes
9	inapp_dose_of_M01AH01 (%)	67 (0.2)	21 (0.2)	18 (0.3)	yes
10	inapp_dose_of_B01AF01 (%)	341 (0.8)	174 (1.7)	115 (1.9)	yes
11	inapp_dose_of_A10BA02 (%)	1452 (3.4)	340 (3.3)	277 (4.6)	yes
12	inapp_dose_of_B01AE07 (%)	257 (0.6)	76 (0.7)	55 (0.9)	yes
13	inapp_dose_of_N05AN01 (%)	68 (0.2)	17 (0.2)	12 (0.2)	yes
14	inapp_dose_of_L04AX03 (%)	19 (0.0)	5 (0.0)	<5	yes
15	inapp_dose_of_B01AF02 (%)	124 (0.3)	184 (1.8)	73 (1.2)	yes
16	n_previous_admissions (%): 0	4988 (11.8)	1082 (10.6)	1074 (18.0)	yes
17	n_previous_admissions (%): 1-2	10100 (23.9)	2367 (23.2)	1873 (31.3)	yes
18	n_previous_admissions (%): 3-4	7712 (18.3)	1919 (18.8)	1232 (20.6)	yes
19	n_previous_admissions (%): 5-6	5490 (13.0)	1303 (12.8)	685 (11.5)	yes
20	n_previous_admissions (%): >6	13960 (33.0)	3530 (34.6)	1116 (18.7)	yes
21	n_distinct_drugs_before_index (%): 0	6165 (14.6)	1228 (12.0)	762 (12.7)	yes
22	n_distinct_drugs_before_index (%): 1-2	9111 (21.6)	1984 (19.4)	1254 (21.0)	yes
23	n_distinct_drugs_before_index (%): 3-4	8761 (20.7)	2078 (20.4)	1355 (22.7)	yes
24	n_distinct_drugs_before_index (%): 5-6	7197 (17.0)	1852 (18.2)	1095 (18.3)	yes
25	n_distinct_drugs_before_index (%): >6	11016 (26.1)	3059 (30.0)	1514 (25.3)	yes
26	admitted_tue (%)	7092 (16.8)	1755 (17.2)	1016 (17.0)	yes
27	admitted_mon (%)	7667 (18.1)	1803 (17.7)	1066 (17.8)	yes
28	admitted_thu (%)	6665 (15.8)	1626 (15.9)	927 (15.5)	yes
29	admitted_sat (%)	3809 (9.0)	960 (9.4)	606 (10.1)	yes
30	admitted_sun (%)	4140 (9.8)	989 (9.7)	598 (10.0)	yes
31	admitted_wed (%)	6718 (15.9)	1616 (15.8)	936 (15.7)	yes
32	admitted_fri (%)	6159 (14.6)	1452 (14.2)	831 (13.9)	yes
33	n_persons (%)	27253 (64.5)	8412 (82.5)	5341 (89.3)	yes
34	n_women (%)	13759 (32.6)	4049 (39.7)	2629 (44.0)	yes
35	n18_diag_any (%)	13470 (31.9)	3391 (33.2)	732 (12.2)	yes
36	icd10_chapter_4 (%)	19716 (46.7)	5096 (50.0)	2415 (40.4)	yes
37	icd10_chapter_9 (%)	25757 (61.0)	6392 (62.7)	3283 (54.9)	yes
38	icd10_chapter_14 (%)	23025 (54.5)	5819 (57.0)	2306 (38.6)	yes
39	icd10_chapter_18 (%)	18663 (44.2)	5711 (56.0)	2882 (48.2)	yes
40	icd10_chapter_19 (%)	20275 (48.0)	4749 (46.6)	2481 (41.5)	yes
41	atc_level2_A02 (%)	10635 (25.2)	2776 (27.2)	1407 (23.5)	yes
42	atc_level2_B01 (%)	11842 (28.0)	3181 (31.2)	1795 (30.0)	yes
43	atc_level2_C03 (%)	13966 (33.1)	3672 (36.0)	1951 (32.6)	yes
44	atc_level2_J01 (%)	14719 (34.8)	3257 (31.9)	1938 (32.4)	yes
45	atc_level2_N02 (%)	15740 (37.3)	4367 (42.8)	2506 (41.9)	yes
46	r (%): 0	38464 (91.0)	9121 (89.4)	5240 (87.6)	no
47	r (%): 1-2	1282 (3.0)	446 (4.4)	391 (6.5)	no
48	r (%): 3-4	670 (1.6)	200 (2.0)	134 (2.2)	no
49	r (%): 5-6	431 (1.0)	116 (1.1)	75 (1.3)	no
50	r (%): >6	1403 (3.3)	318 (3.1)	140 (2.3)	no
51	age_at_admission (median [IQR])	76.9 [66.6, 85.1]	77.4 [68.2, 85.5]	78.5 [69.3, 86.2]	no
52	hour_of_admission_cyclical (%): 0	3560 (8.4)	884 (8.7)	514 (8.6)	no
53	hour_of_admission_cyclical (%): 1-3	18460 (43.7)	4521 (44.3)	2727 (45.6)	no
54	hour_of_admission_cyclical (%): 4-6	12209 (28.9)	2886 (28.3)	1629 (27.2)	no
55	hour_of_admission_cyclical (%): 7-9	5367 (12.7)	1250 (12.3)	712 (11.9)	no
56	hour_of_admission_cyclical (%): 10-12	2654 (6.3)	660 (6.5)	398 (6.7)	no
57	sex_female (%)	20743 (49.1)	4854 (47.6)	2940 (49.2)	yes
58	sex_male (%)	21507 (50.9)	5347 (52.4)	3040 (50.8)	no

59	n18_diag_N189 (%)	13228 (31.3)	3242 (31.8)	694 (11.6)	no
60	n18_diag_N185 (%)	1970 (4.7)	868 (8.5)	62 (1.0)	no
61	n18_diag_N184 (%)	385 (0.9)	223 (2.2)	23 (0.4)	no
62	n18_diag_N183 (%)	161 (0.4)	138 (1.4)	30 (0.5)	no
63	n18_diag_N181 (%)	44 (0.1)	34 (0.3)	<5	no
64	n18_diag_N182 (%)	61 (0.1)	36 (0.4)	9 (0.2)	no
65	elixhauser_score_ahrq (%): 0	10419 (24.7)	2261 (22.2)	1881 (31.5)	no
66	elixhauser_score_ahrq (%): 1-2	1900 (4.5)	432 (4.2)	273 (4.6)	no
67	elixhauser_score_ahrq (%): 3-4	2078 (4.9)	496 (4.9)	356 (6.0)	no
68	elixhauser_score_ahrq (%): 5-6	4725 (11.2)	1092 (10.7)	609 (10.2)	no
69	elixhauser_score_ahrq (%): >6	23128 (54.7)	5920 (58.0)	2861 (47.8)	no
70	icd10_chapter_1 (%)	10263 (24.3)	2662 (26.1)	954 (16.0)	no
71	icd10_chapter_2 (%)	10386 (24.6)	2748 (26.9)	1410 (23.6)	no
72	icd10_chapter_3 (%)	9226 (21.8)	2285 (22.4)	920 (15.4)	no
73	icd10_chapter_5 (%)	7673 (18.2)	1979 (19.4)	1059 (17.7)	no
74	icd10_chapter_6 (%)	7129 (16.9)	1992 (19.5)	992 (16.6)	no
75	icd10_chapter_7 (%)	11452 (27.1)	2901 (28.4)	1495 (25.0)	no
76	icd10_chapter_8 (%)	5392 (12.8)	1203 (11.8)	650 (10.9)	no
77	icd10_chapter_10 (%)	15905 (37.6)	4131 (40.5)	1930 (32.3)	no
78	icd10_chapter_11 (%)	16451 (38.9)	4136 (40.5)	1921 (32.1)	no
79	icd10_chapter_12 (%)	6220 (14.7)	1597 (15.7)	680 (11.4)	no
80	icd10_chapter_13 (%)	16131 (38.2)	4123 (40.4)	2107 (35.2)	no
81	icd10_chapter_15 (%)	159 (0.4)	25 (0.2)	11 (0.2)	no
82	icd10_chapter_16 (%)	38 (0.1)	11 (0.1)	5 (0.1)	no
83	icd10_chapter_17 (%)	1348 (3.2)	312 (3.1)	93 (1.6)	no
84	icd10_chapter_20 (%)	9 (0.0)	<5	<5	no
85	icd10_chapter_21 (%)	37872 (89.6)	9254 (90.7)	5221 (87.3)	no
86	atc_level2_A03 (%)	1819 (4.3)	395 (3.9)	249 (4.2)	no
87	atc_level2_A04 (%)	792 (1.9)	307 (3.0)	200 (3.3)	no
88	atc_level2_A06 (%)	2129 (5.0)	523 (5.1)	256 (4.3)	no
89	atc_level2_A07 (%)	701 (1.7)	185 (1.8)	95 (1.6)	no
90	atc_level2_A09 (%)	22 (0.1)	9 (0.1)	5 (0.1)	no
91	atc_level2_A10 (%)	5270 (12.5)	1500 (14.7)	814 (13.6)	no
92	atc_level2_A11 (%)	4752 (11.2)	1410 (13.8)	558 (9.3)	no
93	atc_level2_A12 (%)	8578 (20.3)	2330 (22.8)	1182 (19.8)	no
94	atc_level2_A16 (%)	6 (0.0)	5 (0.0)	0 (0.0)	no
95	atc_level2_B02 (%)	1271 (3.0)	295 (2.9)	184 (3.1)	no
96	atc_level2_B03 (%)	3513 (8.3)	946 (9.3)	385 (6.4)	no
97	atc_level2_B05 (%)	4884 (11.6)	1643 (16.1)	1074 (18.0)	no
98	atc_level2_C01 (%)	4318 (10.2)	1068 (10.5)	588 (9.8)	no
99	atc_level2_C02 (%)	563 (1.3)	155 (1.5)	45 (0.8)	no
100	atc_level2_C07 (%)	8491 (20.1)	2362 (23.2)	1189 (19.9)	no
101	atc_level2_C08 (%)	4606 (10.9)	1221 (12.0)	607 (10.2)	no
102	atc_level2_C09 (%)	4717 (11.2)	1183 (11.6)	751 (12.6)	no
103	atc_level2_C10 (%)	5698 (13.5)	1479 (14.5)	820 (13.7)	no
104	atc_level2_D01 (%)	23 (0.1)	9 (0.1)	<5	no
105	atc_level2_D07 (%)	29 (0.1)	8 (0.1)	<5	no
106	atc_level2_G03 (%)	83 (0.2)	29 (0.3)	19 (0.3)	no
107	atc_level2_G04 (%)	836 (2.0)	265 (2.6)	150 (2.5)	no
108	atc_level2_H01 (%)	84 (0.2)	15 (0.1)	12 (0.2)	no
109	atc_level2_H02 (%)	3867 (9.2)	1054 (10.3)	575 (9.6)	no
110	atc_level2_H03 (%)	1600 (3.8)	477 (4.7)	267 (4.5)	no
111	atc_level2_H05 (%)	236 (0.6)	62 (0.6)	5 (0.1)	no
112	atc_level2_J02 (%)	286 (0.7)	84 (0.8)	70 (1.2)	no
113	atc_level2_J04 (%)	22 (0.1)	<5	0 (0.0)	no
114	atc_level2_J05 (%)	244 (0.6)	65 (0.6)	36 (0.6)	no
115	atc_level2_J06 (%)	23 (0.1)	7 (0.1)	<5	no
116	atc_level2_L01 (%)	121 (0.3)	27 (0.3)	15 (0.3)	no
117	atc_level2_L02 (%)	166 (0.4)	61 (0.6)	34 (0.6)	no
118	atc_level2_L03 (%)	12 (0.0)	<5	<5	no
119	atc_level2_L04 (%)	964 (2.3)	207 (2.0)	46 (0.8)	no
120	atc_level2_M01 (%)	1628 (3.9)	401 (3.9)	308 (5.2)	no
121	atc_level2_M03 (%)	247 (0.6)	93 (0.9)	62 (1.0)	no
122	atc_level2_M04 (%)	2341 (5.5)	700 (6.9)	276 (4.6)	no
123	atc_level2_M05 (%)	99 (0.2)	24 (0.2)	17 (0.3)	no
124	atc_level2_N01 (%)	378 (0.9)	99 (1.0)	61 (1.0)	no
125	atc_level2_N03 (%)	1917 (4.5)	557 (5.5)	281 (4.7)	no
126	atc_level2_N04 (%)	174 (0.4)	42 (0.4)	22 (0.4)	no

## 127	atc_level2_N05 (%)	5284 (12.5)	1454 (14.3)	823 (13.8)	no
## 128	atc_level2_N06 (%)	3314 (7.8)	843 (8.3)	424 (7.1)	no
## 129	atc_level2_N07 (%)	260 (0.6)	87 (0.9)	42 (0.7)	no
## 130	atc_level2_P01 (%)	1456 (3.4)	268 (2.6)	125 (2.1)	no
## 131	atc_level2_R01 (%)	19 (0.0)	10 (0.1)	5 (0.1)	no
## 132	atc_level2_R03 (%)	627 (1.5)	191 (1.9)	113 (1.9)	no
## 133	atc_level2_R05 (%)	691 (1.6)	113 (1.1)	60 (1.0)	no
## 134	atc_level2_R06 (%)	974 (2.3)	262 (2.6)	96 (1.6)	no
## 135	atc_level2_S01 (%)	169 (0.4)	40 (0.4)	22 (0.4)	no
## 136	atc_level2_V03 (%)	2827 (6.7)	748 (7.3)	279 (4.7)	no
## 137	atc_level2_V08 (%)	37 (0.1)	6 (0.1)	5 (0.1)	no

Supplementary figures

Figure S1: Mapping of admission hour, two alternatives.

Top: Mapping of admission hour (x axis) to unit circle (blue) and distance from midday (red). Bottom: The unit-circle mapping loses the interval nature of the admission hour on the original scale, whereas the distance-to-midday maintains this quality.

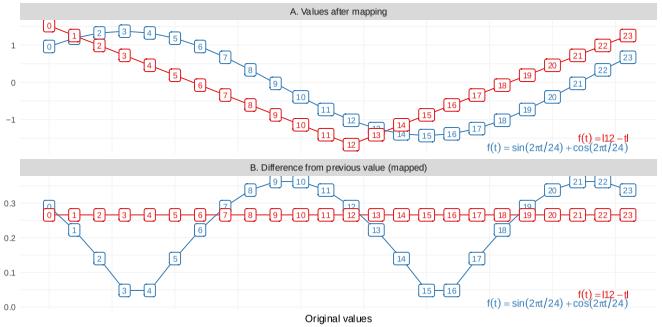


Figure S2: The so-called rulegraph of the our Snakemake pipeline illustrating the end-to-end workflow with dependencies between processing, training and visualisation steps.

Optuna hyperparameter optimisation

Sampled hyperparameter configurations. Point colours represent the loss value: blue = low (preferable), red = high, grey = above the third quartile. Dashed lines indicate the best configuration.

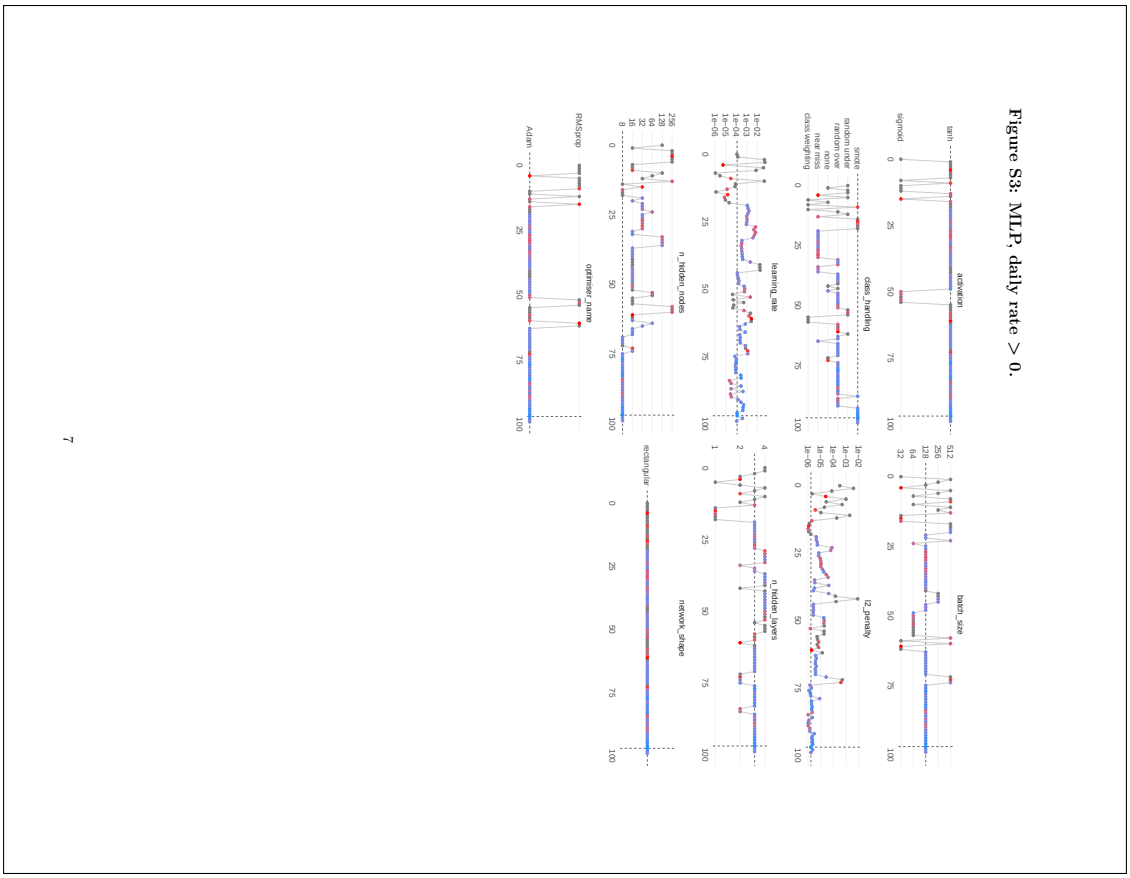


Figure S3: MLP, daily rate > 0.

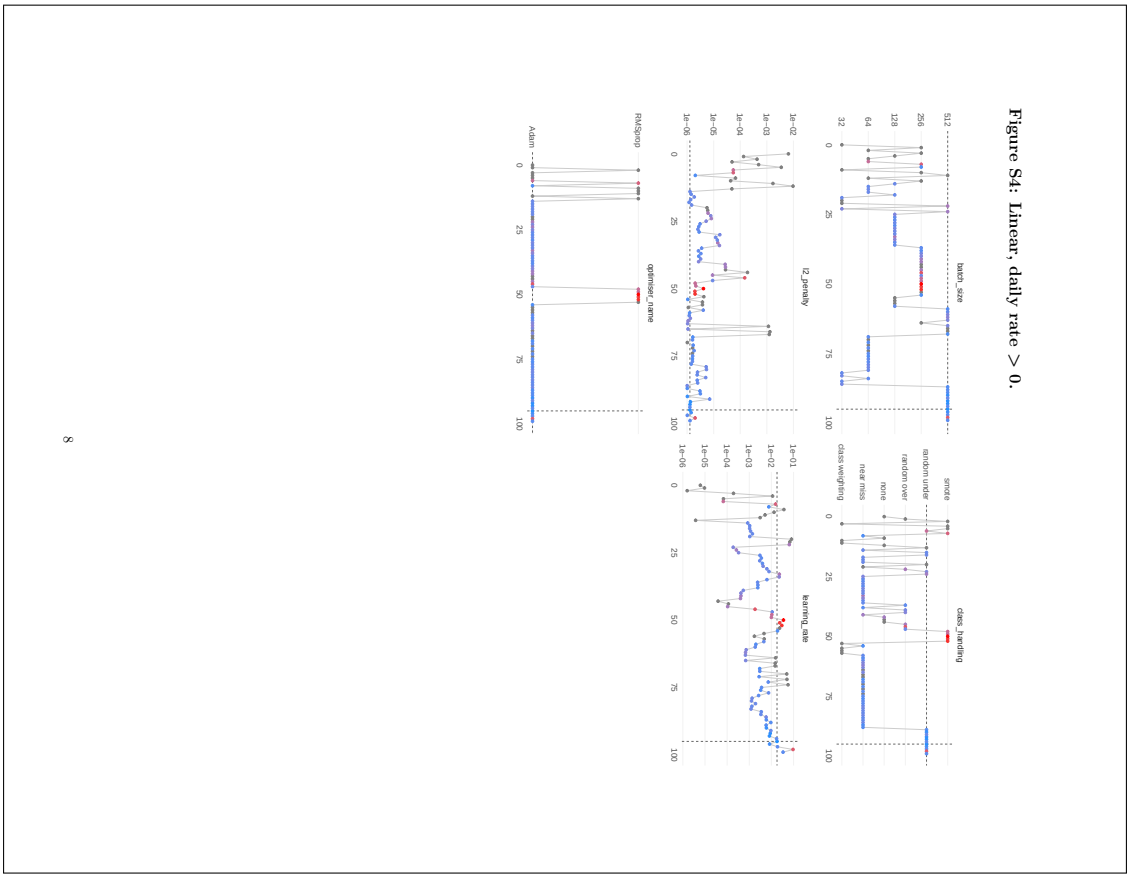


Figure S4: Linear, daily rate > 0.

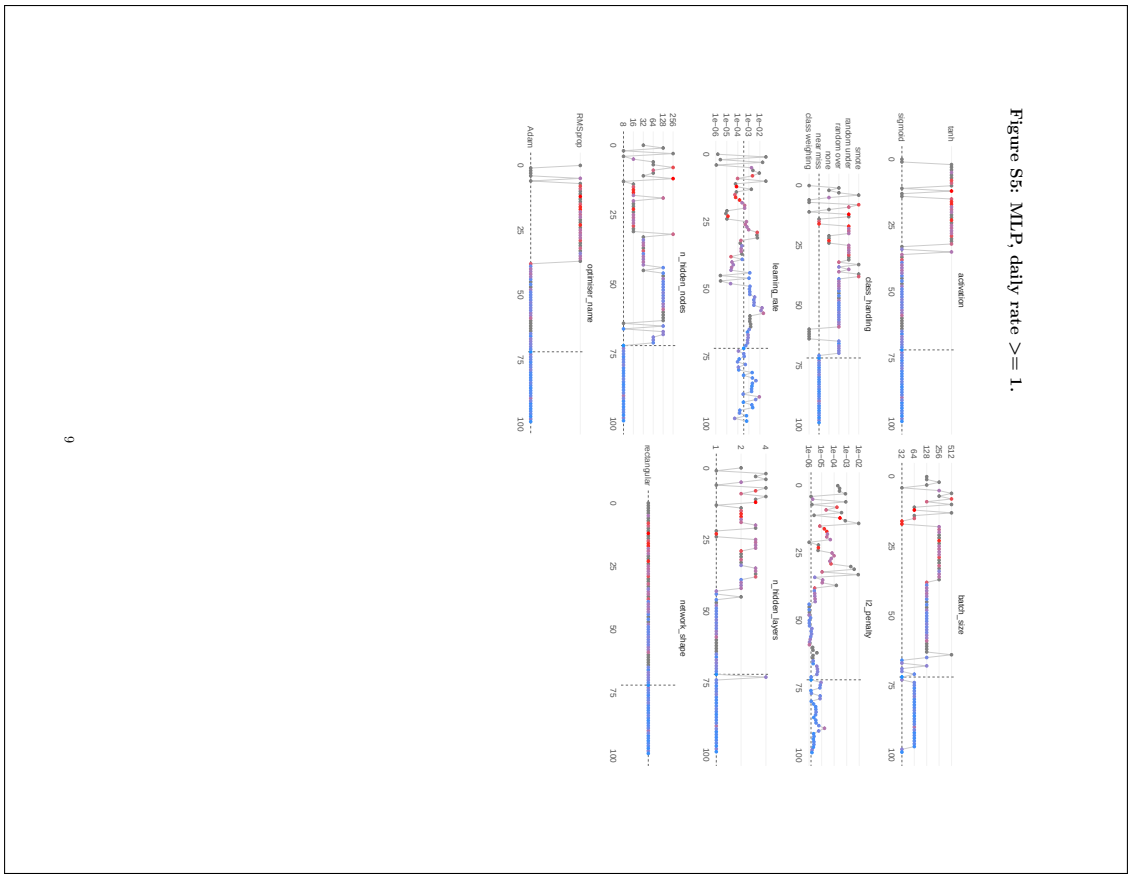


Figure S5: MLP, daily rate >= 1.

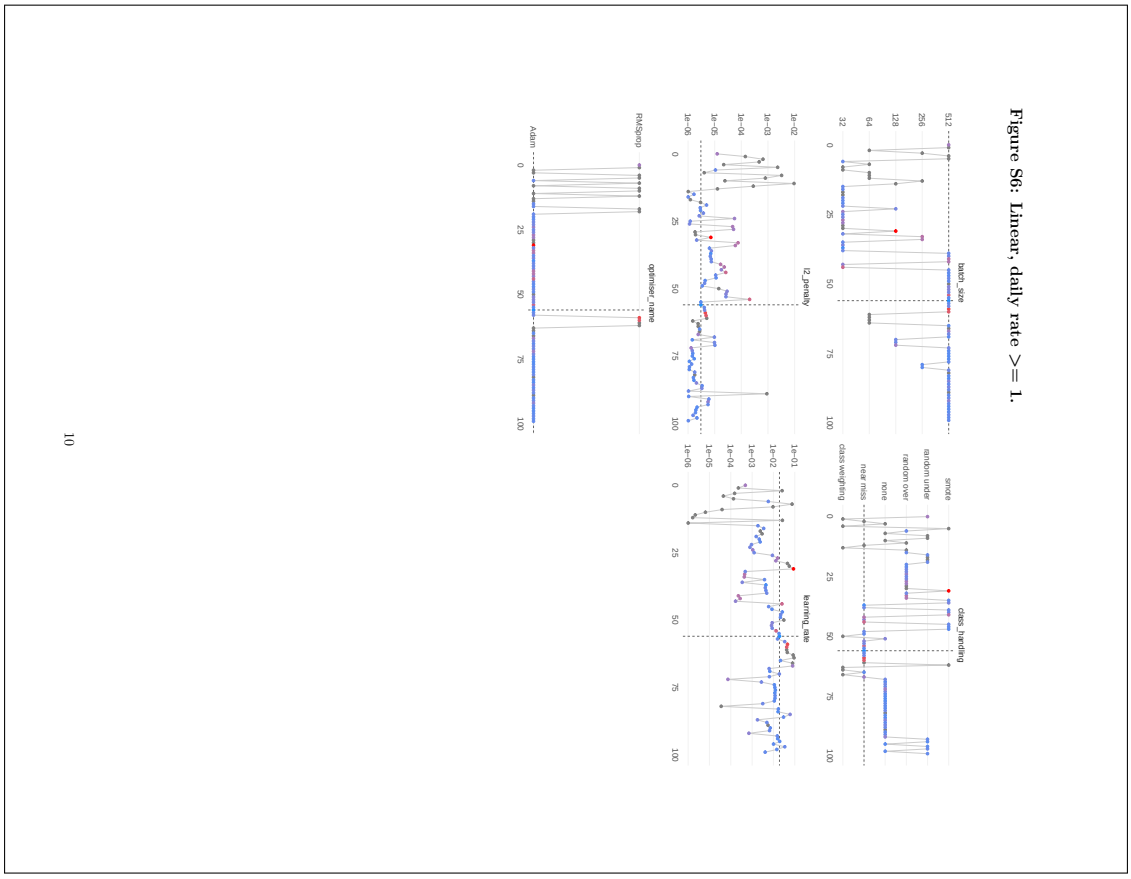


Figure S6: Linear, daily rate >= 1.

Figure S7: MLP, daily rate >= 2.

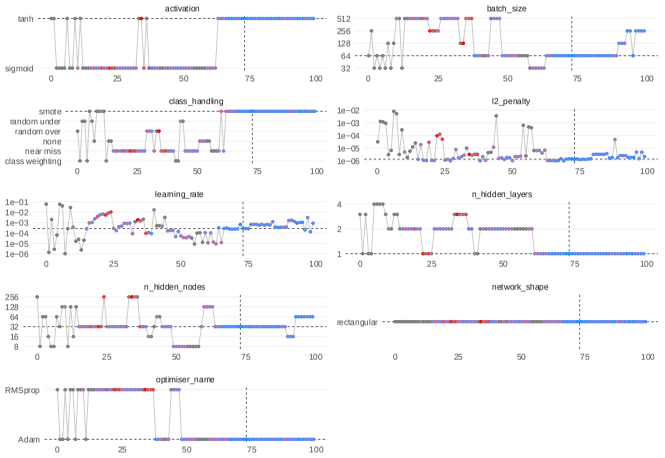


Figure S8: Linear, daily rate >= 2.

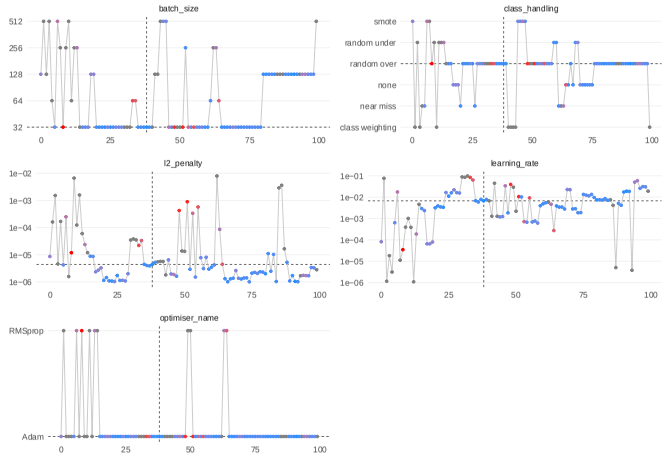


Figure S9: MLP, daily rate >= 3.

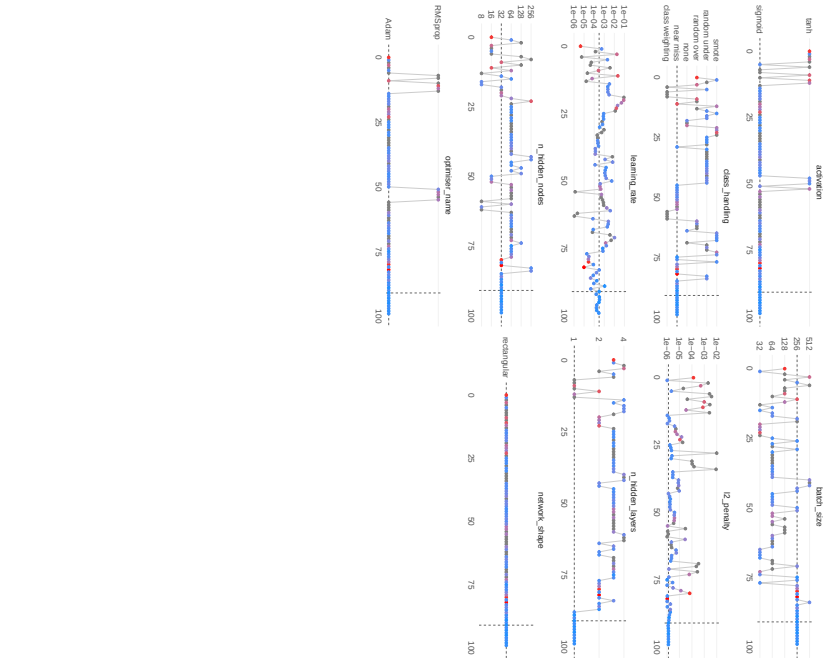


Figure S10: Linear, daily rate >= 3.

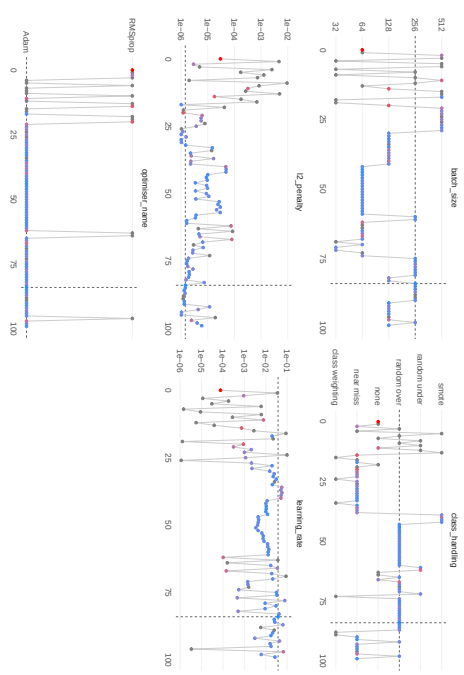


Figure S11: MLP, daily rate >= 5.

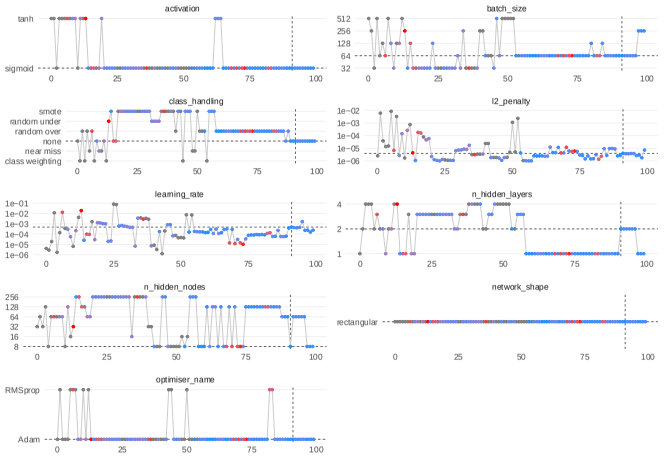
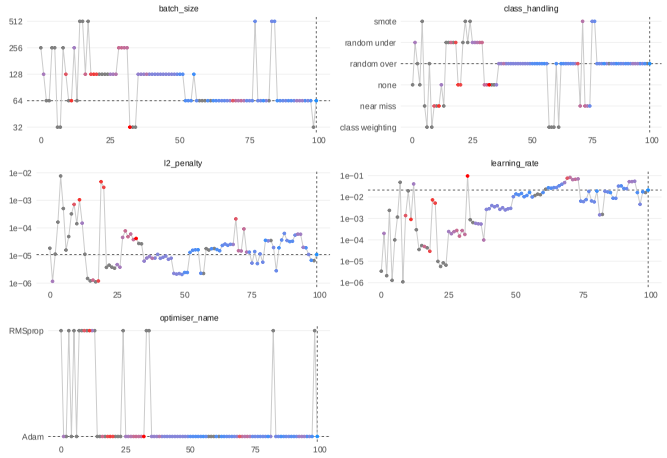


Figure S12: Linear, daily rate >= 5.



ROC curves

Receiver-operating characteristic (ROC) curves in the development and test sets.

Figure S13: MLP, daily rate > 0.

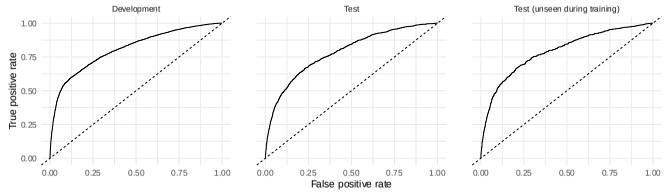


Figure S14: Linear, daily rate > 0.

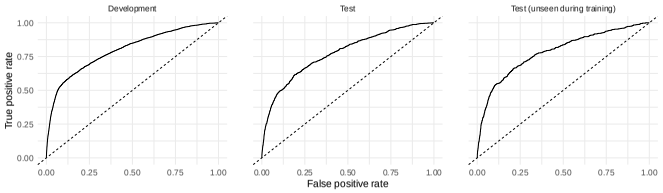


Figure S15: MLP, daily rate >= 1.

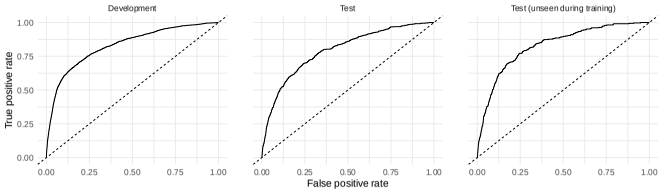


Figure S16: Linear, daily rate >= 1.

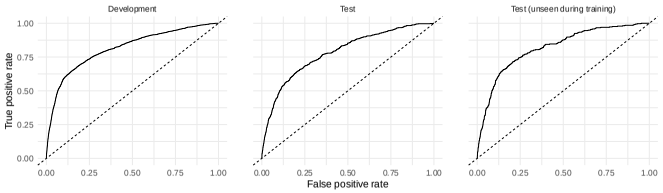


Figure S17: MLP, daily rate >= 2.

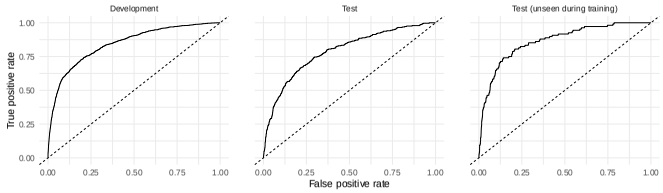


Figure S18: Linear, daily rate >= 2.

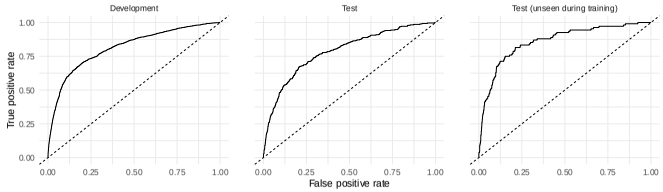


Figure S19: MLP, daily rate >= 3.

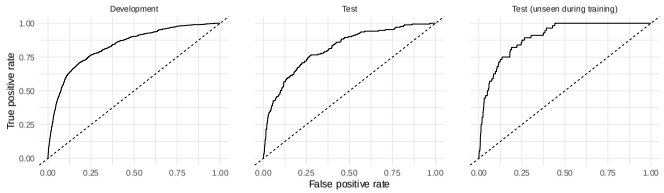


Figure S20: Linear, daily rate >= 3.

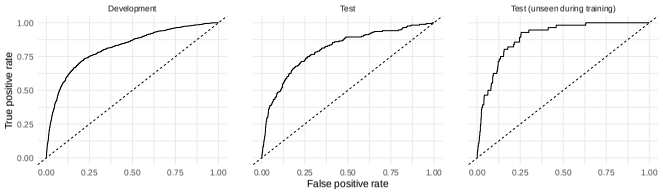


Figure S21: MLP, daily rate >= 5.

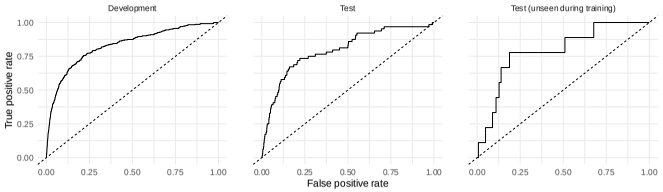
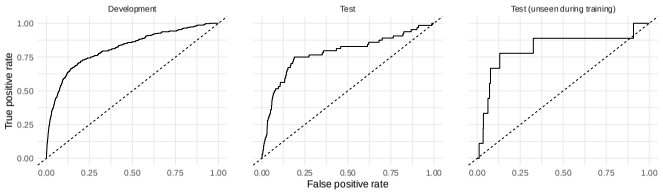


Figure S22: Linear, daily rate >= 5.



Calibration plots

Calibration curves in the development and test sets.

Figure S23: MLP, daily rate > 0.

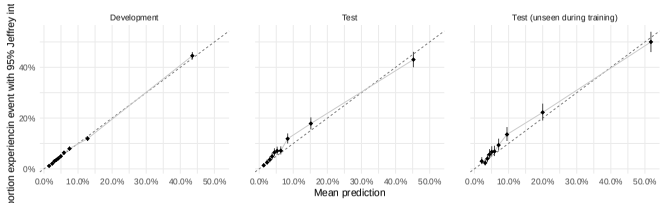


Figure S24: Linear, daily rate > 0.

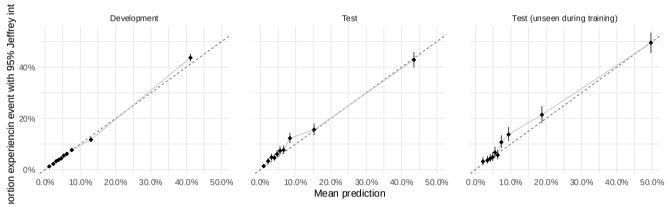
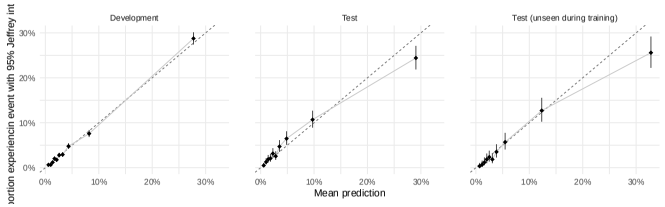


Figure S25: MLP, daily rate >= 1.



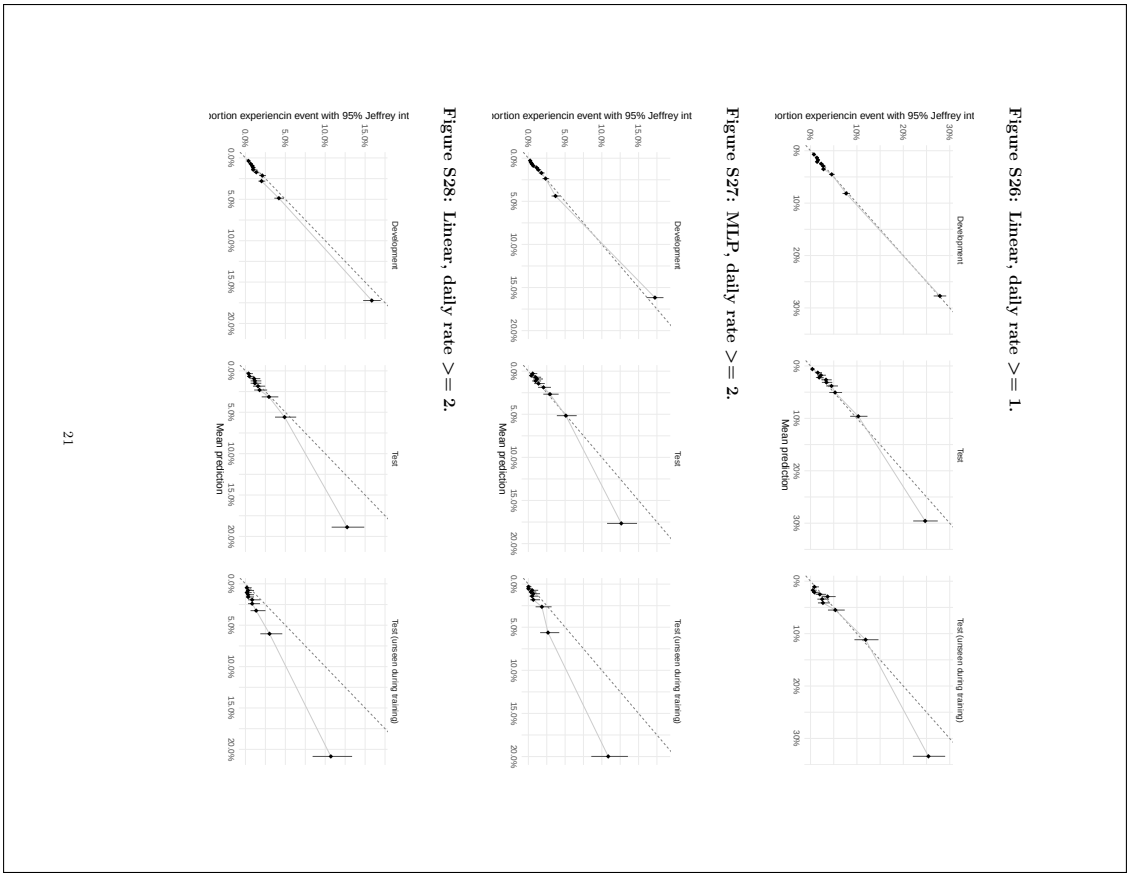


Figure S26: Linear, daily rate >= 1.

Figure S27: MLP, daily rate >= 2.

Figure S28: Linear, daily rate >= 2.

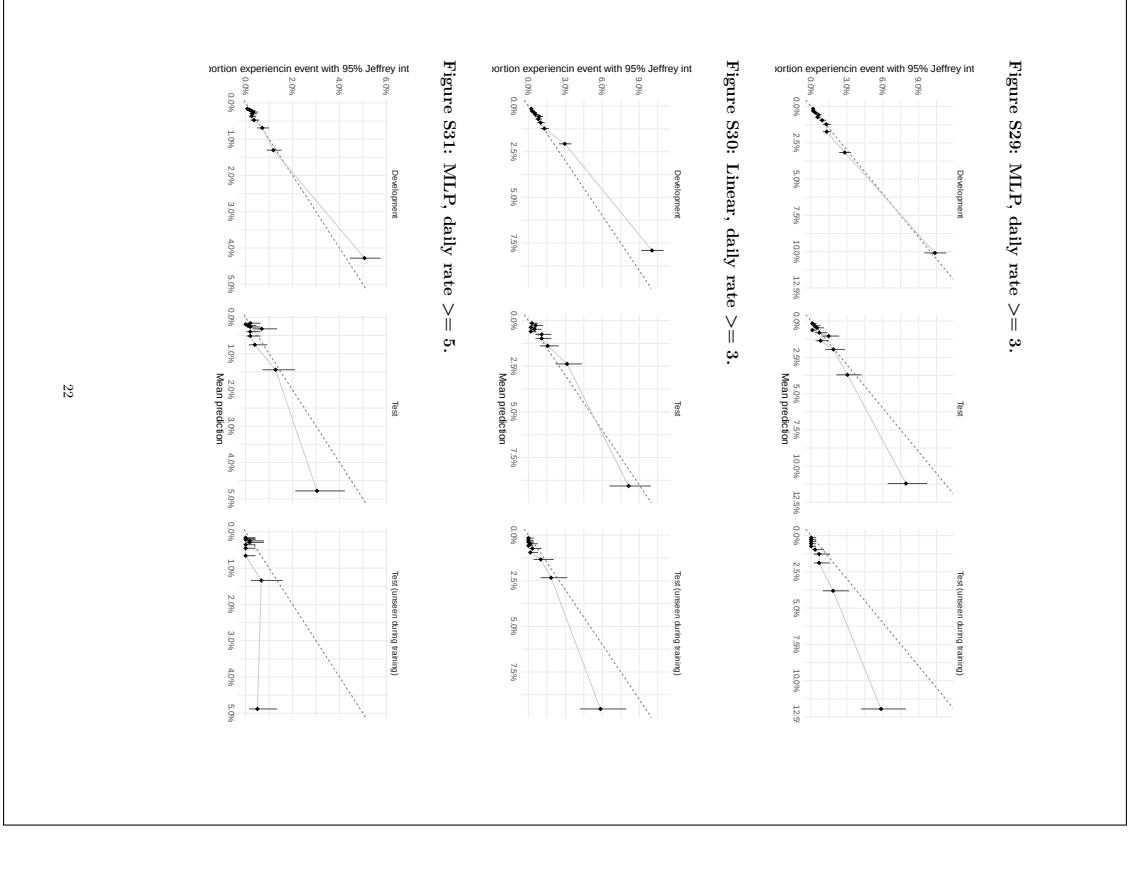
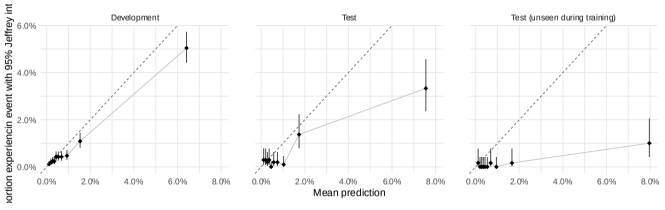


Figure S29: MLP, daily rate >= 3.

Figure S30: Linear, daily rate >= 3.

Figure S31: MLP, daily rate >= 5.

Figure S32: Linear, daily rate >= 5.



Decision curves

Decision curves based on the test set. The curves show the clinical utility (in the unit of standardised net benefit) of intervening in all patients (magenta), no patients (dark grey, follows the x axis) and patients identified by the model (blue).

Figure S33: MLP, daily rate > 0.

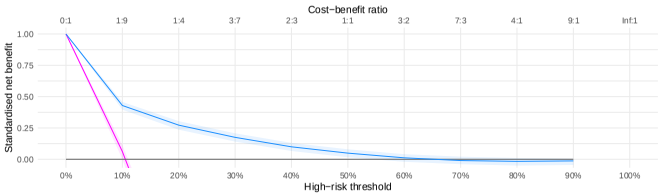


Figure S34: Linear, daily rate > 0.

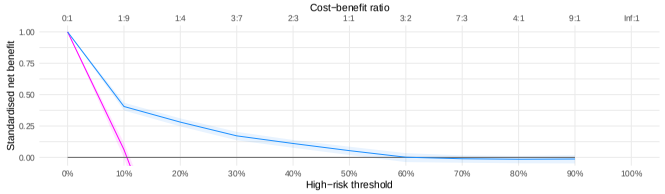


Figure S35: MLP, daily rate >= 1.

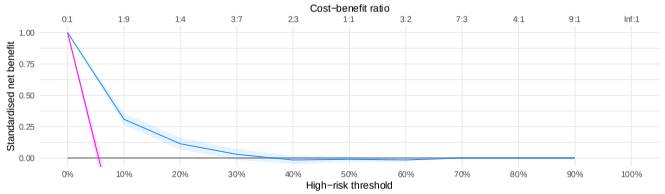


Figure S36: Linear, daily rate >= 1.

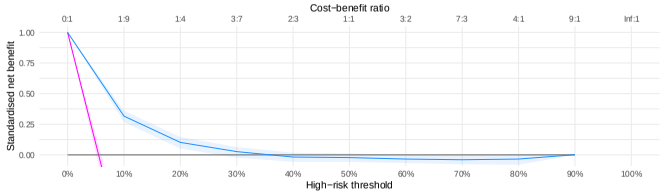


Figure S37: MLP, daily rate >= 2.

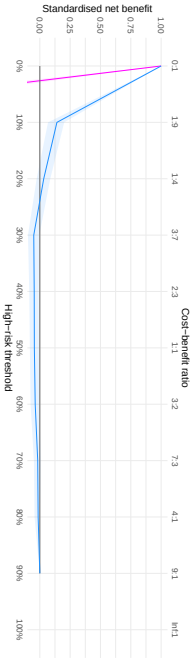


Figure S38: Linear, daily rate >= 2.

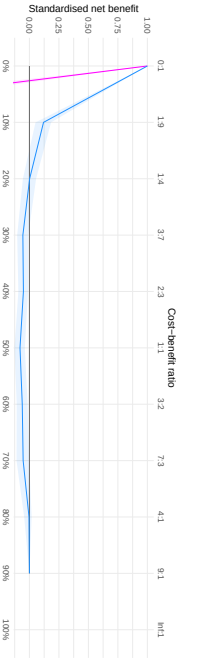


Figure S39: MLP, daily rate >= 3.

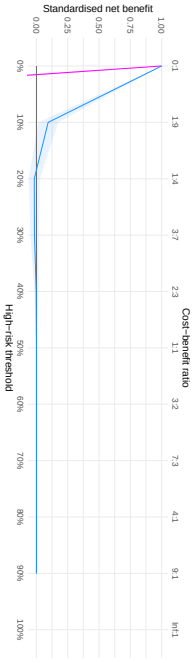


Figure S40: Linear, daily rate >= 3.

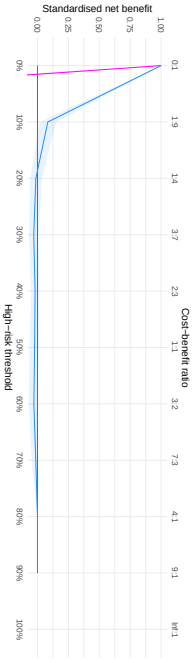


Figure S41: MLP, daily rate >= 5.

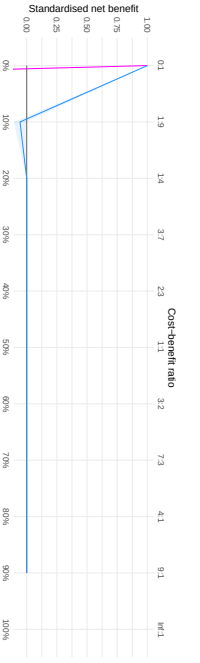
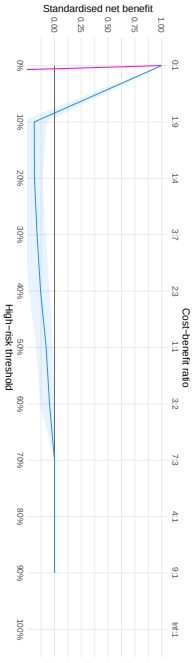


Figure S42: Linear, daily rate >= 5.



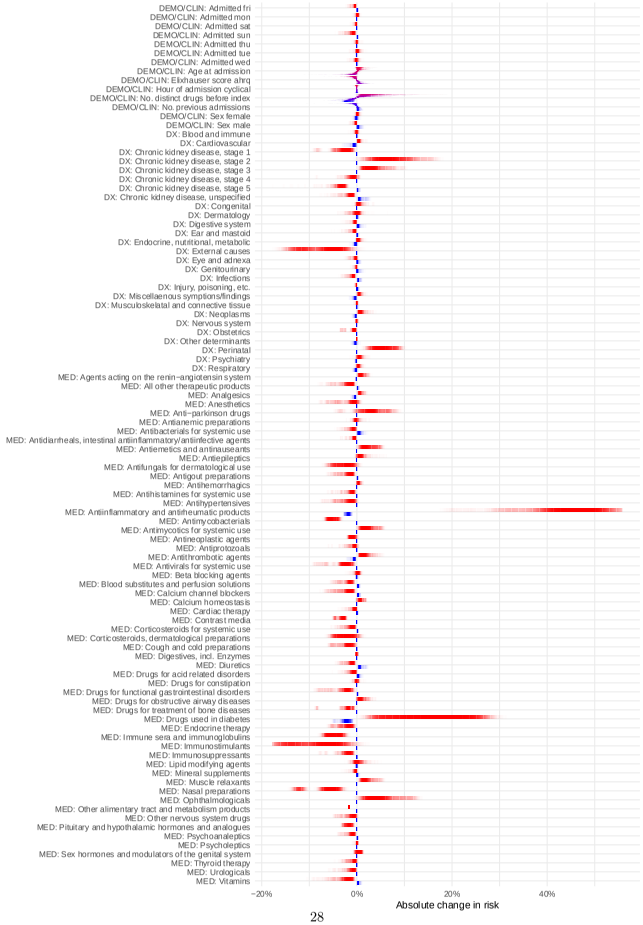
SHAP plots

The individual shap plots (figures S44-53) below all visualise the the distributions as density bands of shap values by feature. Blue represents low feature values (0 for binary features) and red high feature values (1 for binary features). Continuous features were binned into deciles.

Figure S43: Summary plot of shap values across studies

Feature values were binned into at most 5 bins, and each bins is represented by one point: the x axis is the mean shap value for each bin, the colour illustrates the spectrum of feature values (blue = low, red = high). Points are connected by lines to aid reading, solid and dashes lines represent MLP and linear models.

Figure S43: MLP, daily rate > 0



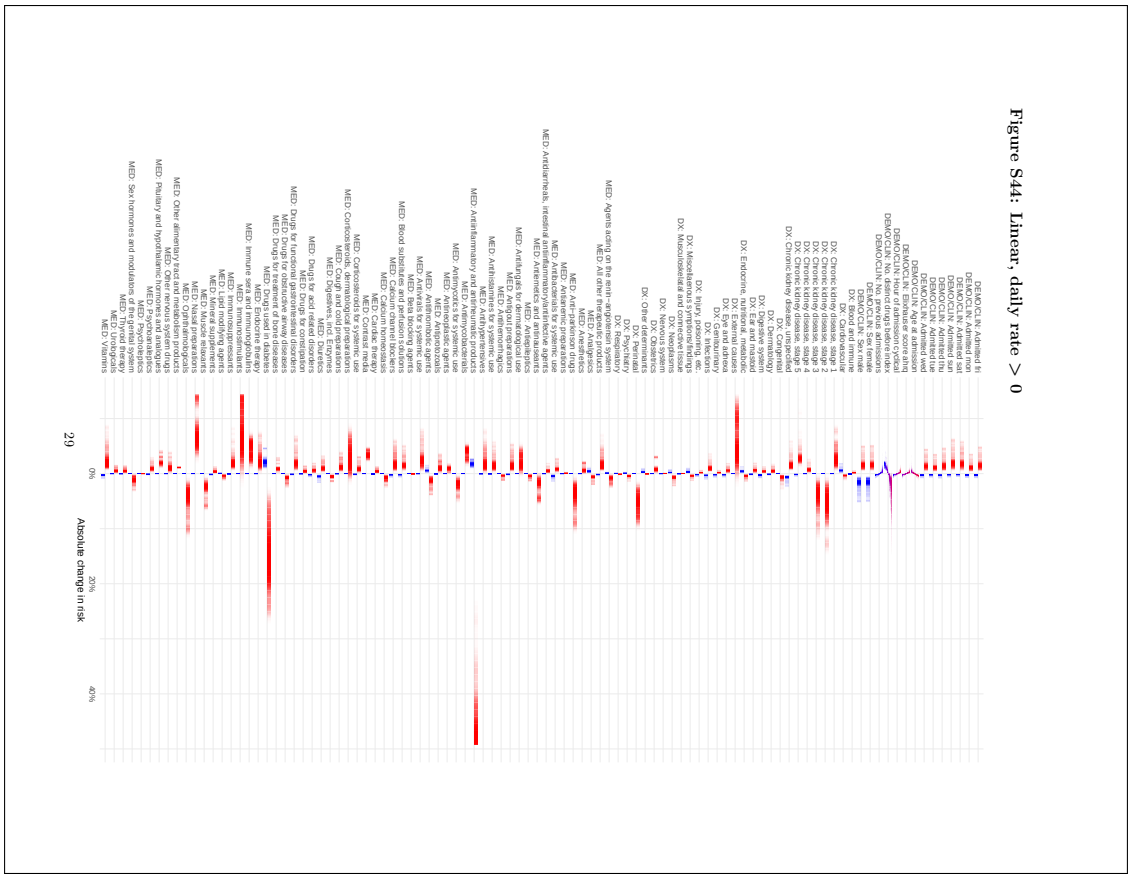


Figure S44: Linear, daily rate > 0

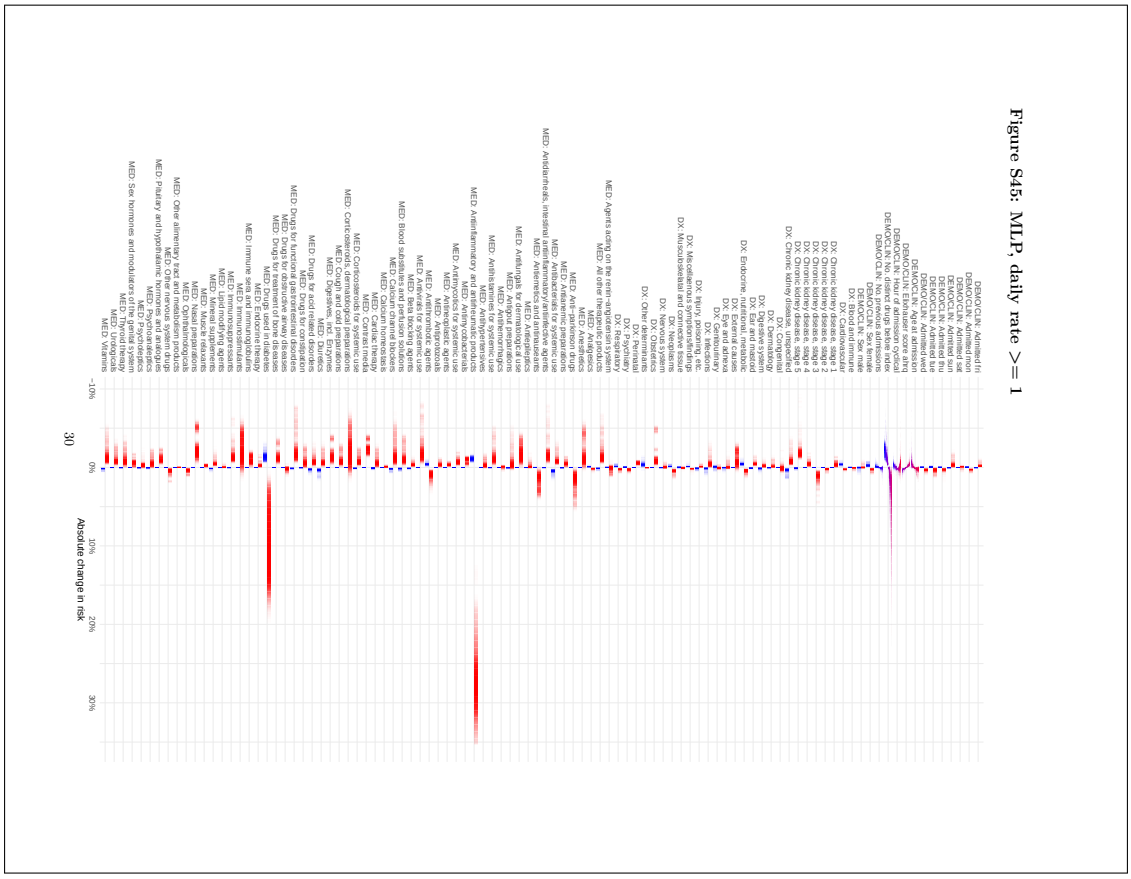


Figure S45: MLP, daily rate >= 1

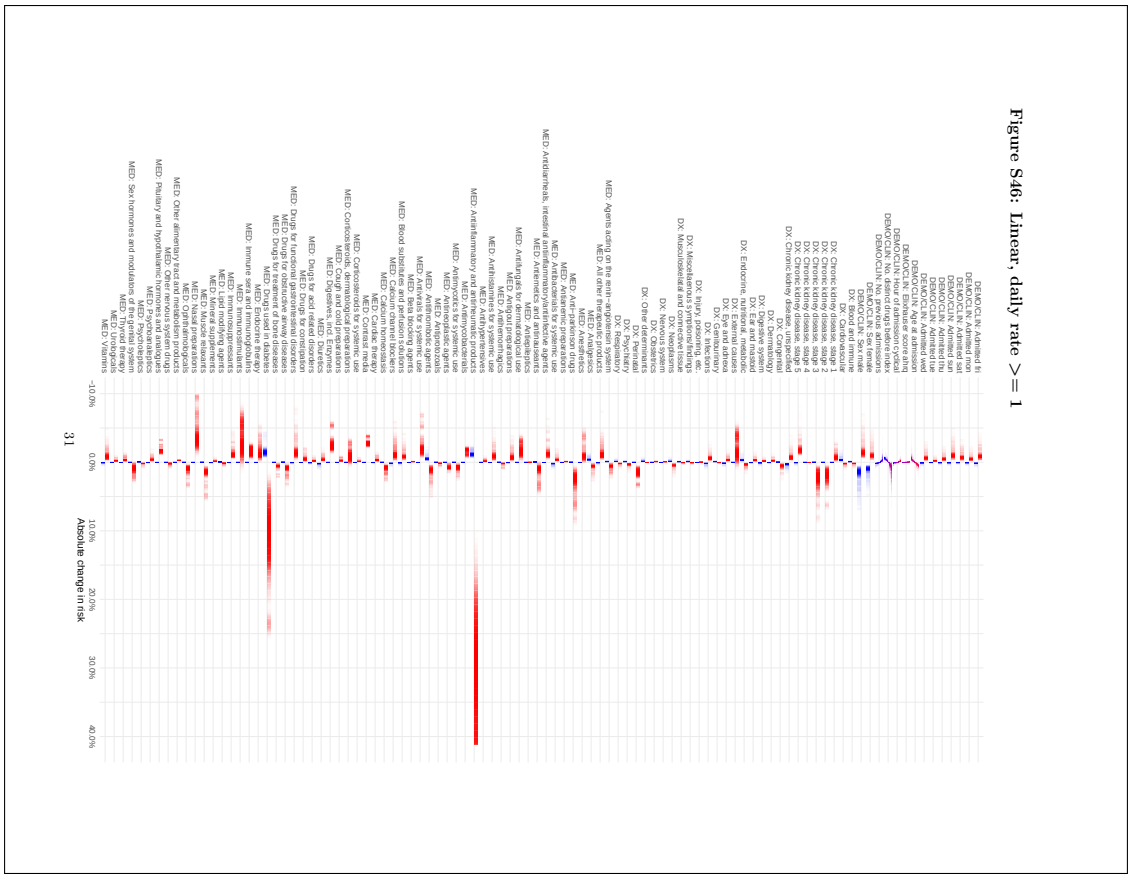


Figure S46: Linear, daily rate >= 1

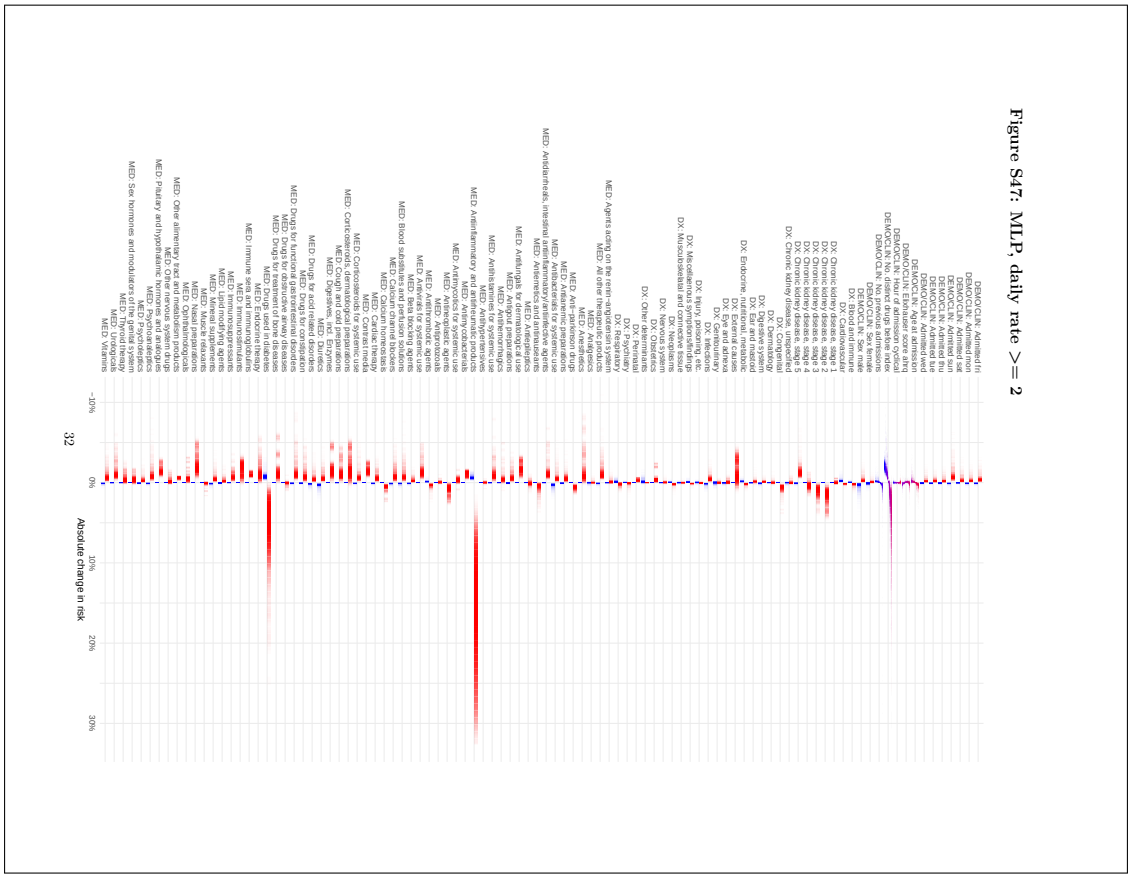


Figure S47: MLP, daily rate >= 2

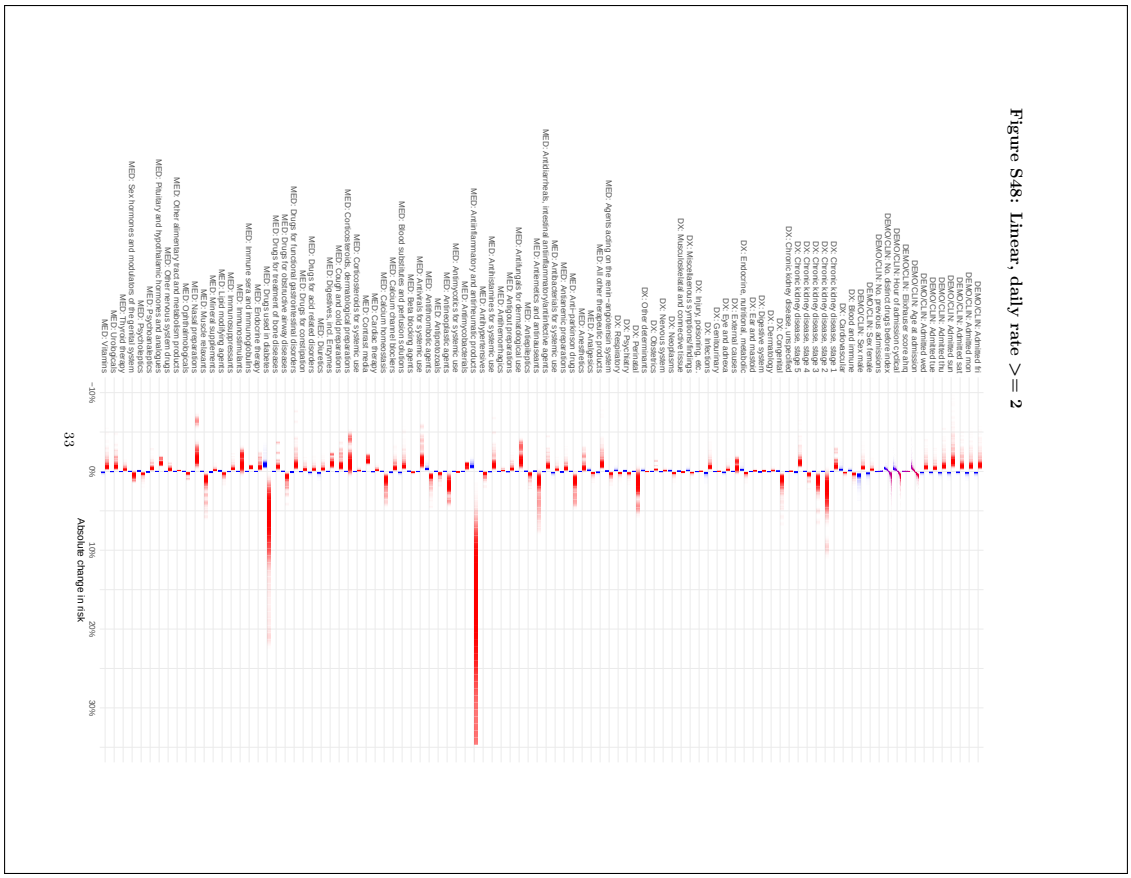


Figure S48: Linear, daily rate >= 2

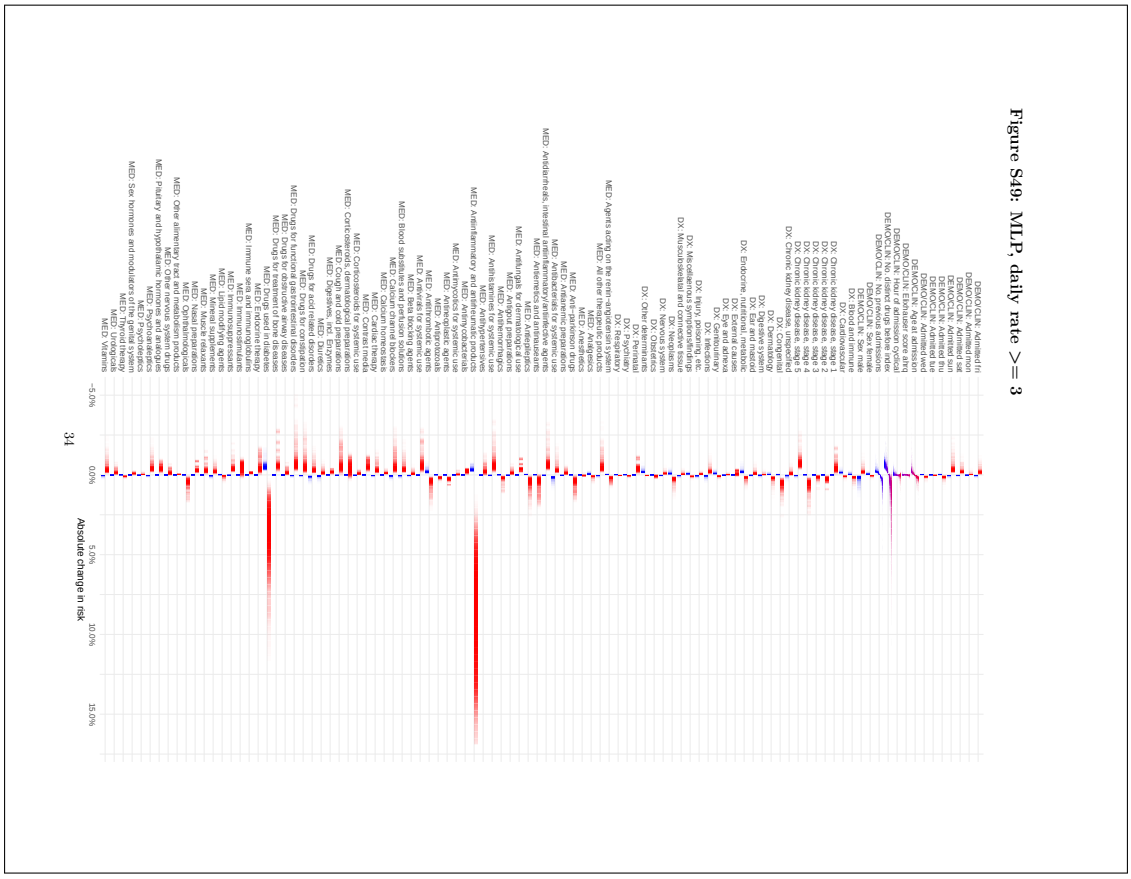


Figure S49: MLP, daily rate >= 3

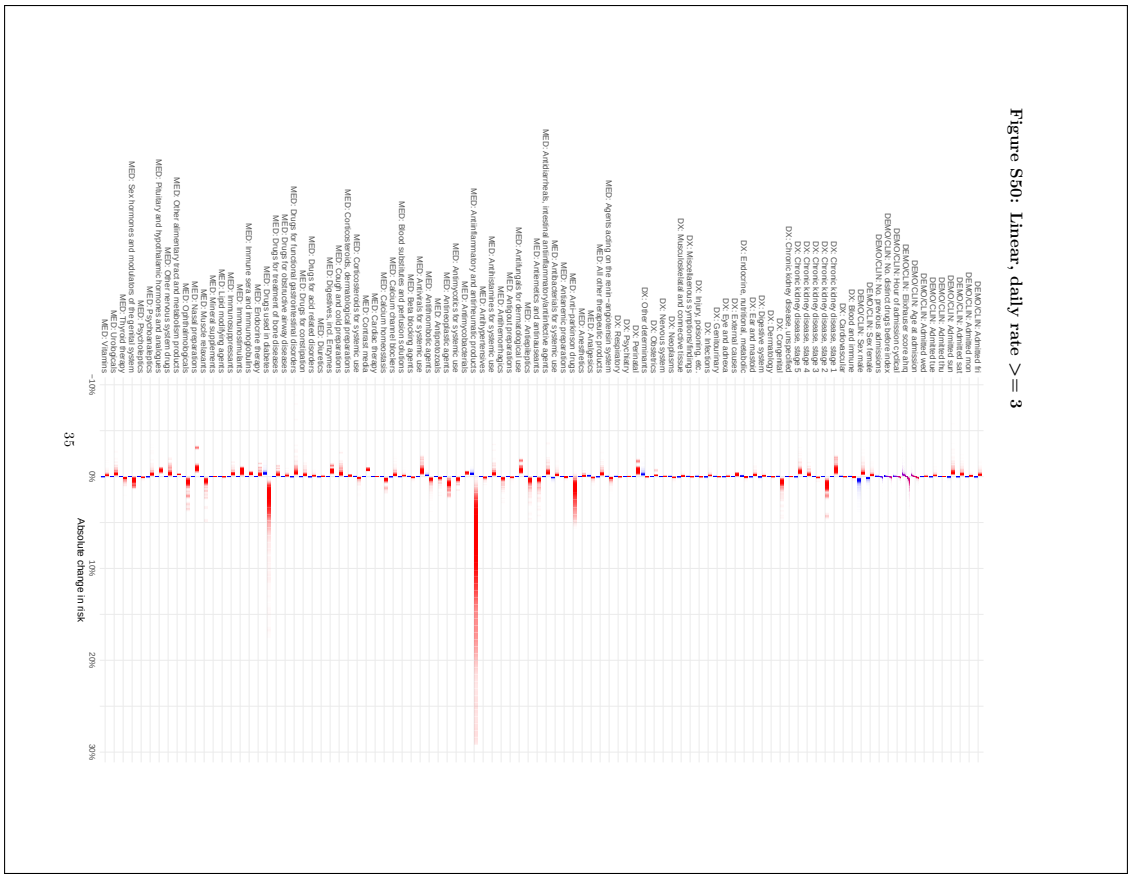


Figure S50: Linear, daily rate >= 3

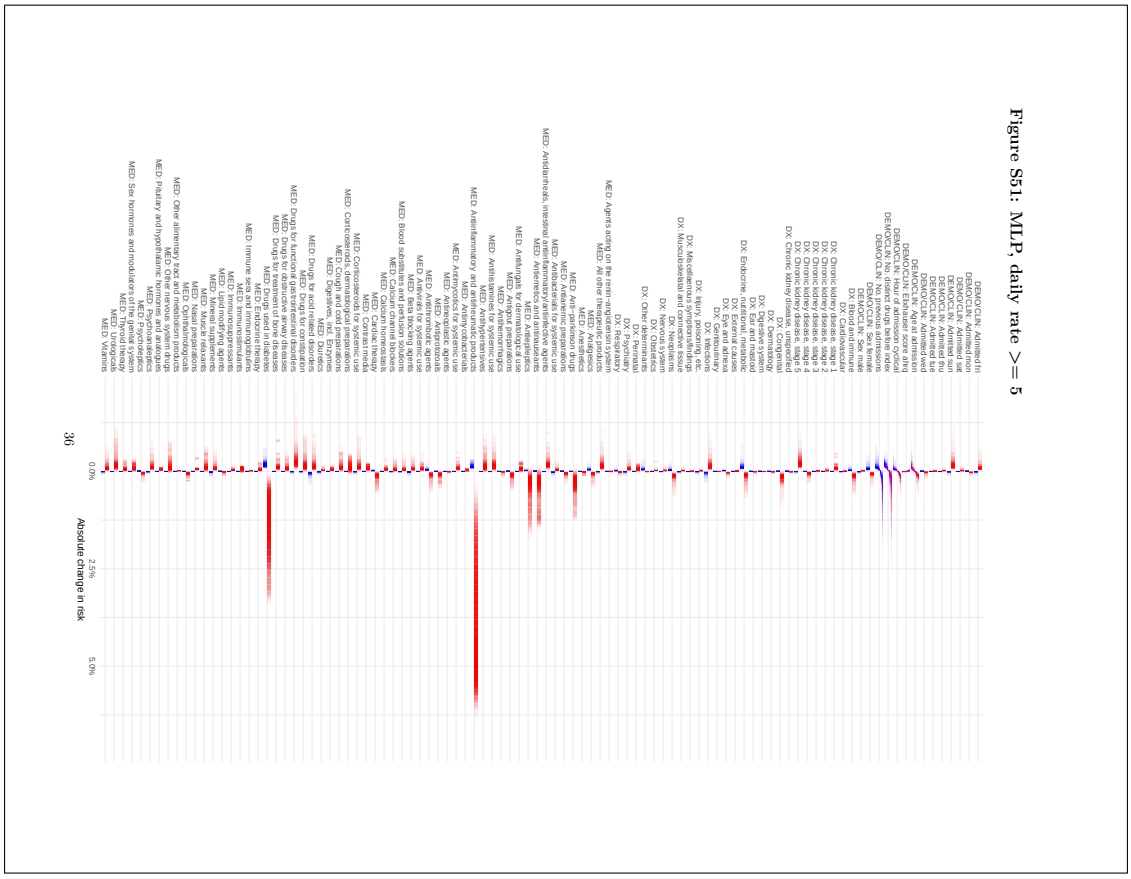


Figure S51: MLP, daily rate >= 5

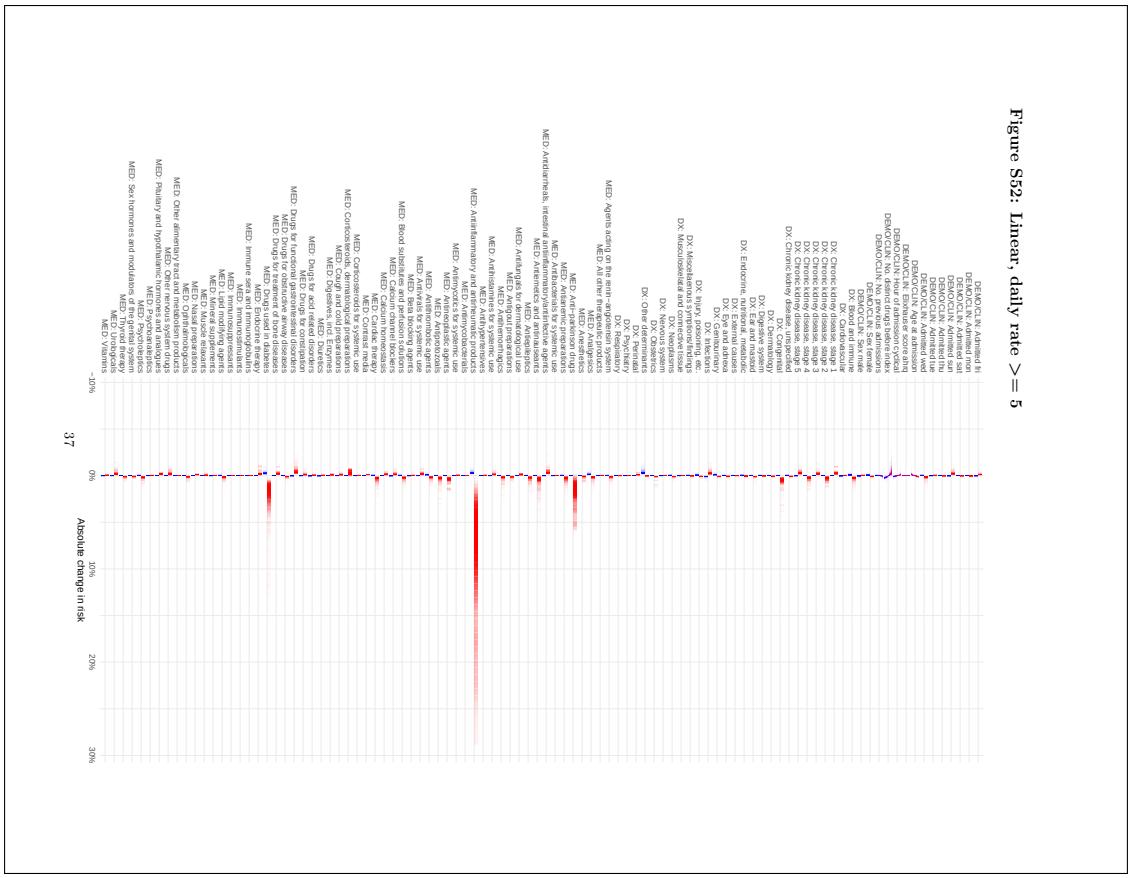


Figure S52: Linear, daily rate $>= 5$

Language-agnostic signal detection in clinical notes

Full title

Eliciting side effects from clinical notes: language-agnostic pharmacovigilant text mining

Chapter contents

Eliciting side effects from clinical notes: language-agnostic pharmacovigilant text mining

Benjamin Skov Kaas-Hansen[1,2,3] MD, MSc — Davide Placido[2] MSc
Cristina Leal Rodríguez[2] MSc, PhD — Hans-Christian Thorsen-Meyer[4] MD, PhD
Simona Gentile[5] MD — Anna Pors Nielsen[2] MD
Søren Brunak[2] MSc, PhD — Gesche Jürgens[1] MD, PhD — Stig Ejdrup Andersen[1] MD, PhD

Affiliations

[1] Clinical Pharmacology Unit, Zealand University Hospital, Denmark

[2] NNF Center for Protein Research, University of Copenhagen, Denmark

[3] Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

[4] Department of Intensive Care Medicine, Copenhagen University Hospital (Rigshospitalet), Denmark

[5] Region Zealand, Denmark

Contributions

Based on the CRediT contributor roles taxonomy; authors listed alphabetically. Conceptualisation: BSKH. Literature review: BSKH, SG. Data curation: BSKH, CLR, DP. Formal analysis: BSKH, CLR, DP. Methodology: BSKH, CLR, DP, HCTM. Software: BSKH, CLR, DP, HCTM. Interpretation: APN, BSKH, HCTM, GJ, SEA, SG. Supervision: GJ, SEA. Funding acquisition: SB, SEA. Resources: SB, SEA. Drafting: BSKH. Review: All.

Conflicts of interest

SB reports ownerships in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S, and managing board memberships in Proscion A/S and Intomics A/S outside the submitted work. All other authors report no competing interests.

Abstract

Purpose

To create a drug safety signalling pipeline associating latent information in clinical free text with exposure profiles to highlight potential adverse drug reactions to single drugs and drug pairs.

Data and methods

All inpatient visits of a 500,000-patient sample from two Danish regions, between 18 May 2008 and 30 June 2016. Tokens from clinical notes recorded within 48 hours of admission were extracted and operationalised with a 256-dimensional fastText embedding; single-drug and drug-pair exposures from doorstep medication profiles were cast into one-hot encoded vectors. For each of the resultant unique 10,720 target exposure we trained a multilayer perceptron with two hidden layers of 256 nodes, predicting the risk of exposure using tokens' embedding vectors as inputs. Only signals from well-calibrated models with good discrimination were considered pertinent for further evaluation: congruence between signals for terms with very similar meaning but different spelling, and manual review by three assessors.

Results

In the included 2,905,251 inpatient visits (1,559,685 (54%) women) the median age was 58 (inter-quartile range, IQR: 33-73) and stable throughout. There were 13,740,564 doorstep drug prescriptions; the median number of prescriptions was 5 (IQR: 3-9) and in 1,184,340 (41%) admissions patients used ≥5 drugs concurrently. 10,788,259 clinical notes were included, with 179,441,739 tokens (per-admission median: 51 [IQR: 29-80]) retained after pruning. 3,945 (38%) models yielded pertinent signals.

Congruence was good: signal profiles agreed within UKU terms, within UKU domains, and within the mental-neurological spectrum. Inter-rater agreement was moderate. Of 345 single-drug signals reviewed, 28 (8.1%) represented possibly undescribed relationships; for 186 (54%) signals, the reactions were possible, known, or due to protopathic or indication bias, all clinically meaningful relationships. 16 (14%) of the 115 drug-pair signals were possible interactions and 2 (1.7%) were known.

Conclusion

We have successfully built a language-agnostic pipeline for mining associations between free-text information and medication exposure without the need for manual curation. We achieve this by turning things upside down, predicting not the likely outcome of a range of exposures, but the likely exposures for one or several outcomes of interest. Our

approach may help overcome limitations of text mining methods relying on curated data in English. This makes our method appealing in settings that must make sense of non-English free text for pharmacovigilance while, with few adaptations, lending itself well to alternative use cases such as patient-level decision-making support and drug repurposing.

Introduction

Pharmacovigilance usually operates with two qualifications of the common term *side effect*: adverse drug events (ADEs) and adverse drug reactions (ADRs). ADEs are (noxious) medical events occurring while using certain medicines without assuming causal relationships between [1]. ADRs are subsumed by ADEs and constitute outcomes believed or known to be *caused* by exposure to a given medicinal product [2,3][2,3]. ADRs are usually classified in 6 groups, including *dose-related* and *not dose-related* [4] The latter are more unpredictable than the former and tend to be unrelated to the pharmacological effect, making them interesting from a safety signal detection perspective.

ADRs signal detection usually revolves around spontaneous case reports, collated nationally (e.g. Danish Medicines Agency), regionally (e.g. European Medicines Agency) and internationally (e.g. VigiBase of the Uppsala Monitoring Centre [5]). This system suffers from several shortcomings, including the inherit filtering of reports making it into central databases, causing i.a. under-reporting [6-9] that may even be biased or otherwise influenced by, for example, media hype or legislation [10] although the number of spontaneous reports from biopharmaceutical companies is growing [11]. These weaknesses, and the ever-expanding digitisation of patient data, have sparked much interest in leveraging complimentary data sources and technologies for pharmacovigilance, including longitudinal clinical data and natural language processing (NLP), the branch of machine learning for making textual data compatible statistical modelling [12,13].

Text mining is a subfield of NLP dedicated to extracting structured information from inherently unstructured textual data. Text mining applications in pharmacovigilance often hinge on hand-curated reference sets for named-entity recognition or entity extraction [14-17]; for example, previous work brought about a Danish dictionary of side effects [18]. These tasks focus on assigning labels to free-text terms so they can be codified and used as structured data akin to diagnostic codes recorded in national registers [19] or adverse-event databases.

Creation and maintenance of such gold standards are costly and tedious, which likely explains the limited availability of tools and resources (including corpora) for non-English textual data. For example, the official ADR vocabulary of the Danish Medicines Agency is MedDRA (with English terms), and submitters of case reports are encouraged to pick from English terms when submitting case reports. When non-standard side effects are entered, these are manually mapped to the English MedDRA afterwards. Thus, although many a

little makes a mickle, it is near-impossible to extract information across languages which would be useful for pharmacovigilant purposes. We posit that, to leverage clinical free text, complementing existing vocabulary-based approaches to pharmacovigilant NLP with unsupervised and automatic information extraction from clinical free text deserve exploration and could facilitate vast screening of clinical free text.

To this end, we report on the creation of one such complementary system: an end-to-end machine learning pipeline associating latent information in clinical free text with medication profiles to highlight potential adverse drug reactions to single drugs and two-way drug combinations. We envision a system that accepts one of several side-effect terms from the user and returns likely, prominent exposures that would undergo assessments akin to the evaluation of signals in spontaneous case reports.

Methods and materials

Data were obtained from electronic patient records (EPR) systems of 12 secondary and tertiary public hospitals in two Danish regions (Capital Region and Region Zealand), comprising approximately 2.6 million persons (about half the Danish population). We used data from 500,000 adult (age ≥ 18 years) patients admitted between January 2006 and 30 June 2016.

The full analytic workflow is depicted schematically in figure 1 and has four main components (detailed below): doorstep medication profiles (red), embedding model (brown), operationalisation of clinical notes (blue), training the signal detection component (green), and evaluating the safety signals (purple).

Doorstep medication profiles

We considered only pre-existing medication at start of admission and created one medication vector with one element per distinct single drug and drug pair in the full data set, using their respective anatomical therapeutic chemical (ATC) codes. Elements corresponding to drugs and drug pairs used by that patient at doorstep were set to 1, the rest to 0. We only considered single drugs and drug pairs used in at least 1,000 admissions.

Embedding model

An embedding packs high-dimensional data into much fewer dimensions. Imagine, for example, one-hot-encoding words in a corpus of clinical notes that collectively contain 345,671 unique words; the presence of a word in a given note could be represented by a (very sparse) vector with 345,670 zeros and a single 1. Learning a 100-dimensional embedding of the words, in contrast, enables us to represent each word by a 100-element

vector that captures latent information in unstructured text [13]. This vector will not be sparse (computationally convenient) and vectors of words with similar meaning will be similar even when lexicographically different (e.g. *headache*, *sore head* and *neuralgia*).

We used fastText [20] to train the embedding model on the full corpus after slight pruning: characters other than letters and numbers were removed, as were multiple white spaces. This yielded one white-space separated string of words from each note. Hyperparameters were arbitrary but appropriate for the task at hand; for example, we used a 256-dimensional embedding, sub-word components were allowed to be between 3 and 6 characters long (*minn* and *maxn* settings), and tokens were allowed to span up to 3 words to capture multi-word signals (such as *chest pain* or *sore head*; *wordNgrams* setting). All settings can be found in the analytic code, see below.

Operationalisation of clinical notes

The corpus comprised notes recorded within the first 48 hours of admission; each note underwent five processing steps. First, the note was split into sentences. Second, within each sentence we identified negations and for each of these excluded the subsequent 5 words or until end-of-sentence (heuristic based on Thomas et al. [21].) Third, we removed special characters from these non-negated words. Fourth, we retained the pruned words that were neither Danish stop words (using *nltk.corpus* [22]) nor present in an in-house list of almost 430,000 names used in Denmark. We forewent stemming and lemmatisation to let the model learn from natural words, to facilitate its downstream use. Finally, these retained tokens were concatenated by admission, essentially considering each admission one document (an oft-used term in text-mining and information retrieval literature).

We computed the term-frequency/inverse-document-frequency (TF-IDF) as $tf \times log(N/(1 + df))$ for each retained token (with $10 \leq df \leq 50{,}000$) to automatically filter away tokens so common or rare that they unlikely contained information of interest [23]. The final TF-IDF values were not used to discard tokens at this step; that happened during training, see below.

The final step of this component was converting tokens to their corresponding embedding vectors using the fastText model, which happened while training to not unnecessarily store vectors for tokens many of which were never used due to under-sampling, see below.

Training the signal detection component

We constructed one multilayer perceptron (MLP) model with 2 hidden layers of 256 nodes for each of the 12,270 unique drugs and drug pairs in the medication profiles, setting the

binary outcome to 1 if that drug (pair) was in the doorstep medication profile and 0 otherwise. Because of the imbalanced nature of the prediction task (figure 2) and to obtain tolerable runtime, we used random 1:2 under-sampling of the majority class to help the model focus on pertinent signals. We used all tokens for cases and the top-50 tokens based on TF-IDF for controls. We, then, used the embedding vector for each token and used that with its outcome as one observation in the MLP model.

We used sigmoid activation functions, the Adam optimiser and regularisation only in the form of early stopping based on area under the receiver operating characteristic curve (AUROC) in the internal validation set. The validations set came about by 80/20 random split-sampling, deemed appropriate as this served solely for regularisation and not validation per se [24].

Pertinence was operationalised as signals from well-performing models with respect to discrimination and calibration-in-the-small using the 20% internal validation set. Discrimination was gauged by AUROCs, calibration-in the-small by the intercepts and slopes of linear regressions to the calibration curves of decile-binned predicted probabilities and corresponding bin-wise observed outcome probabilities [25]. Only models with intercepts in [-0.05, 0.05], slopes in [0.95, 1.05] and AUROCs ≥ 0.7 in the validation sets were considered to yield pertinent signals.

Evaluating safety signals

Congruence

To quantify the relevance of the signals we compared the predicted odds with the odds in the background population and used these *odds ratios* as the signal scores.

The congruence analysis served to qualitatively assess whether tokens with near-identical or very similar clinical meanings ("clinical synonyms") were assigned the same medication profiles regardless of lexicographical similarity or lack thereof. To this end, we used the terms in figure 5 (see also next section for details on their origin) and a list of clinical synonyms for a total 116 terms. Congruence was, then, assessed visually by plotting pairwise adjusted cosine distances [26] between the signal profiles of all 116 terms, constructed as the union of all exposures in the top-50 of any of the terms.

Relevance

We used a reference set to gauge the signals' relevance, that is to what extent signals are meaningful from a clinical and pharmacovigilance point of view. Several potential reference sets exist [27], but for three principal reasons chose the items in the UKU (Udvalg for Kliniske Undersøgelser, English: Committee for Clinical Investigations) side

effect rating scale [28]. First, the UKU items were originally developed in a Nordic setting, so the English-Danish translations are readily available. Second, the UKU items were developed to gauge the side-effect load of psychotropics, and so their (somewhat) well-defined pharmacological mechanisms aid the assessment of biological plausibility of signals (more on this below). Third, our results are readily put in a scientific context because the UKU scale has been used for several years and in different contexts [29], ensuring transparency with respect to and confidence in the translations for readers unfamiliar with the Danish language.

We manually reviewed the top-5 single-drug and top-5 drug-pair signals for each reference-set term consulting three standard sources in clinical pharmacology: www.pro.medicin.dk (side effects; identical side-effect information as the official Danish summaries of product characteristics (SPCs, available at www.produktresume.dk) with few exceptions), DrugBank (drug-drug interactions; publicly available information; www.drugbank.ca [30]) and the Danish Interaction Database [31] (drug-drug interactions). We crafted a helper R package (promedreadr, doi: 10.5281/zenodo.5529817) to do the heavy lifting when collecting side-effect information from www.pro.medicin.dk. DrugBank kindly made their data (v5.1.8) available to the first author for the purpose of this study.

Each single-drug signal was labelled as (a) example of protopathic bias or bias-by-indication [32], (b) known side effect if reported for at least one product with that ATC code, (c) possible side effect (i.e. biologically plausible), or (d) spurious signal (in this order). For drug-pair signals we labelled each drug according to the single-drug classification and further evaluated the signal from a drug-drug interaction point of view on two axes: whether the two drugs are known to interact (is any interaction described in the Danish Interaction Database and/or DrugBank?) and relevance of signal (three options: known result of interaction, possible result of interaction, or not caused by interaction). BSKH, GJ and SEA undertook signal assessment: each signal was evaluated independently by two assessors and disagreement (quantified by Cohen's kappa [33]) was resolved by consensus.

Ethics

This study is part of the BigTempHealth research programme for which approval was granted by the Danish Patient Safety Authority (3-3013-1723; then competent authority for ethical approval), the Danish Data Protection Agency (DT SUND 2016-48, 2016-50, 2017-57) and the Danish Health Data Authority (FSEID 00003724). This report honours the RECORD statement [34] as relevant.

This study's codebase is available online (doi: 10.5281/zenodo.5598068). Trained on the full corpus, the embedding model contains sensitive information and so cannot be shared with third parties.

Results

The final data set covered the period from 18 May 2008 through 30 June 2016 and comprised 2,905,251 inpatient visits (admissions) of which 1,559,685 (54%) were of women. The median age was 58 (inter-quartile range, IQR: 33-73) and stable throughout the study period. These admissions comprised 10,788,259 clinical notes (18% of these patients' 60,960,247 notes) recorded within 48 hours of admission and 13,740,564 doorstep drug prescriptions; the median number of prescriptions in the doorstep profiles was 5 (IQR: 3-9) and in 1,184,340 (41%) admissions patients used ≥5 drugs concurrently, a common polypharmacy threshold [35]. Pruning and filtering left 179,441,739 tokens (per-admission median: 51 [IQR: 29-80]) for training the 10,270 neural-network models of which 3,945 (38%) yielded pertinent signals.

Figure 2 shows the relative frequency of all 571 single-drug exposures and (correspondingly) the top-571 drug-pair exposures. The dominant drug classes were those affecting the nervous system (N, including psychiatric drugs), the alimentary tract and metabolism (A), and the cardiovascular system (C). The same picture emerged from the drug-pair exposures: the most prevalent drug pairs involved these same three drug classes (e.g. AA, AC and AN).

We devised so-called fingerprints for each main UKU term visualising single-drug exposures (figure 3). These fingerprint plots illustrate that general or vague terms (e.g. depression, nausea, weight gain) are relatively strongly associated with many drug exposures, and that for more specific terms (e.g. amenorrhoea, galactorrhoea) fewer drugs, of appropriate drug classes, light up. Also, fingerprints of clinically related terms (e.g. tremor, parkinsonism and dystonia) are similar but clearly distinct from those of other terms.

Congruence

We hypothesised that signal profiles would be similar for similar side-effect terms ("clinical synonyms"), regardless of lexicographical similarity or lack thereof. Indeed, as figure 4 illustrates, signal profiles agreed within UKU terms, within UKU domains, and within the mental-neurological spectrum. As expected, the terms in the *Other* domain did not agree well, likely because this domain comprises very different side effects not fitting

in elsewhere. Agreement was imperfect, which can be seen from e.g. the light stripes representing terms with signal profiles distinct from all other terms. Several UKU terms have synonyms identical to those of other UKU terms so these of course will show perfect congruence, even if across UKU domains.

Relevance

Agreement between the three assessors (BSKH, GJ, SEA) was moderate, with four values of Cohen's kappa (κ): relevance of drug 1 (κ = 0.49), relevance of drug 2 (κ = 0.72), whether the two drugs were known to interact in any way (κ = 1.0) and relevance of interaction (κ = 0.73). The consensus assessments in figure 5 clearly shows that the method picked up pertinent information.

There were 345 single-drug/potential-reaction pairs (figure 5, caption). Of these, 28 (8.1%) represented possible relationships between drug exposure and the reaction in question (figure 5B, light green). For 186 (54%) signals the reactions were either possible, known, or due to protopathic or indication bias, all clinically meaningful relationships (figure 5B, green and dark grey). 16 (14%) of the 115 drug-pair signals were possible interactions, 2 (1.7%) were known and the rest not attributable to the drugs interacting (figure 5C).

Discussion

With a novel, language-agnostic approach using word embeddings we successfully built an end-to-end pipeline to elicit potential side effects of out-of-hospital drug exposure. The method may complement existing signal screening processes through automated detection of possible side-effects. Using side-effects from the psychiatric domain with (somewhat) well-defined pharmacological properties we illustrated that this method may offer genuine utility: manual review of signals for clinically relevant side effects illustrated the ability of the pipeline to highlight pertinent signals, with the "hit rate" in the same order of magnitude as that of signal detection in spontaneous case reports [36].

The novelty of our approach hinders direct comparisons with the published literature. Indeed, we try to fill a gap in the three-axis categorisation of pharmacovigilance NLP: using non-English text, removing the reliance on annotated data, and leveraging EHR data. The number of published NLP applications in pharmacovigilance is growing: a review from 2012 included but 7 studies, most of which used either simplistic keyword searches or more elaborate NLP methodologies (MediClass, MedLEE), predominantly in discharge summaries with relatively old data (1995 through 2008) [37]. More recently, a review from 2017 included 48 studies and emphasised the need for side-effect detection

methods to handle also polypharmacy-related side-effects [38], an issue intimately related to drug-drug interactions.

Side-effect signal detection generally occurs in three types of data (spontaneous case reports, online forums—including social media—and longitudinal patient data) with the analytical approaches somewhere along two axes (modelling complexity and structuredness of the data). The long-standing signal detection in spontaneous case reports rests on several large database (e.g. FAERS, EudraVigilance and VigiBase) collecting reports from healthcare staff, patients and pharmaceutical companies across the globe. The mainstay of this system has been disproportionality analytic [39] with attempts at assessing DDIs [40], although NLP applications exist [41-44] just as several attempts at leveraging online content for pharmacovigilance have come about [45-49], especially using Twitter posts [50-59] with examples of trying to disentangle temporality of exposure-event pairs [60].

Although pharmacovigilant text mining in non-English corpora is not the norm, examples do exist. A Danish dictionary of side effects was created and used for mining psychiatric patient files, relying on ontologies against which terms found in the clinical text were compared [18,61,62] and, thus, different in scope than ours. Oronoz et al. sought to create a gold standard from EMR notes in Spanish that had been annotated by pharmacologists and pharmacists, with particular focus on medicines and diagnoses [63], while Segura-Bedmar et al. sought to extract drug effects, both beneficial and noxious, from a Spanish online health forum [64]. Another study used Japanese online platforms to evaluate basic characteristics of medicine users [65] and Ujiie et al. used medical articles, manually annotated by a medical engineer, in Japanese articles published for postmarketing surveillance [66]. Usui et al. devised a system to automatically assign ICD-10 codes to Japanese free-text patient complaints recorded by pharmacists when dispensing prescription medicines [67].

These examples all share the foundational characteristic that they rely on curated ontologies for annotating their corpora. This eases evaluation as the curation process establishes a ground truth against which to compare the algorithm's output. Nevertheless, real-life clinical corpora are moving targets, and the constant expansion and morphing of ontologies require continual and costly updating of annotation rules. Our approach stands in contrast to this: it is an end-to-end pipeline that requires no annotation of specific documents but acts a simple signal detection engine whose signals should then undergo expert review. With text embedding at its core, the method allows for data augmentation [68] without hand-tuning; we did not, however, venture down this path.

Data mining models generally carry no causal meaning, and an oft-raised issue of NLP is the need for (often large) annotated corpora which requires much work and continuous updating to remain relevant, the very thing we attempted to circumvent by reversing the prediction direction. Others have used word embeddings to operationalise free text in a non-annotated manner. For example, Workman et al. showed that word embeddings can help overcome the problems of misspelling in a pharmacovigilance application [69]; the RedMed model was trained on Reddit posts to extract *health entities* therein and performed reasonably well in such consumer-generated content [70]; and combining pre-trained word embeddings and conditional random fields could have flagged potential cutaneous adverse reactions to two chemotherapy classes in internet content before they were reported in the scientific literature [48].

We trained one model per drug exposure for a total of 10,270 individual models. Although multi-label architectures sometimes aid learning [71], we found this to drown pertinent signals in models with thousands of outputs nodes in a single network. This probably happens because the model can only optimise a single loss value and we found no good way to automatically up- or down-weigh contributions from different outputs. Further, in a multi-label feed-forward architecture all weights but those between the last hidden layer and the outputs are shared, and there seems to be no good reason that predicting the risk of, say, exposure to metformin should be so intimately linked to that of olanzapine.

When designing our approach, we had institutional pharmacovigilance efforts in mind, but alternative use cases exist, such as patient-level decision-making support and drug repurposing research. Including patient characteristics (e.g. age, sex and comorbidities) would enable clinical staff to query the method for single drugs or drug combinations potentially explaining the symptoms of their patients. Instead of looking at drugs given disproportionately often for a given term, we could focus on those given more rarely (so, with the odds ratio < 1) potentially eliciting interesting novel target conditions for existing treatments similar in spirit to e.g. Kessing et al. [72].

Combinatorial explosion is a well-known challenge for the study of DDIs: a person using 7 different medicines is exposed to 21 two-way drug combinations. This challenge is only exacerbated if higher-order combinations are considered. So, instead of modelling this explicitly one could consider higher-order interactions (e.g. 3- or 4-way) by piecing together two-way combinations that yield predicted probabilities above a certain threshold when multiplied, i.e. using a simplistic approximation to the predicted joint probability.

An alternative approach, and indeed research question, would have been to compare new in-hospital exposures with terms in subsequent days for immediate side effects. To be feasible this would likely require a much larger data set to have sufficient exposure-outcome pairs. It might, however, be less unwieldy as such an approach could focus on new(er) drugs drastically reducing the number of labels (and, thus, models to be trained).

Strengths and limitations

Our approach has five principal strengths. First, its unsupervised nature drastically reduces the need for manual work. This sets it apart from most other published studies using NLP in pharmacovigilance that tend to hinge on manual curation. Second, the method is language-agnostic owing to its unsupervised nature, so that it does not rely on a vocabulary to look up words. This renders the approach potentially useful for pharmacovigilance in also smaller languages and machine translation could enable screening efforts in disparate textual data sources of different languages even by people with no knowledge of the language(s) in question. Third, our corpus is quite large, a natural consequence of its non-reliance on curated data. Fourth, skipgrams (i.e. using sub-word information) enable embedding of also i.a. word bigrams, misspellings, and out-of-vocabulary words. Fifth, the crude and almost reductionist nature of our approach circumvents many difficulties posed by NLP because we break documents down to basic components and use them without modelling semantics and syntax.

This study, however, is subject to several limitations. First, the apparently well-defined temporality obtained using doorstep medication profiles does not necessarily guarantee that what is reported in the text occurred after start of exposure. This potential problem, also the source of protopathic bias [32], is not unique to our approach but rather necessitates cautious interpretation of any signal detection method, in longitudinal and case-report settings alike. Second, we do not actually have data on prescriptions from the primary sector but rely on the doorstep registration of pre-existing medication. Physicians are obliged to record these doorstep medication profiles, and we expect they generally be accurate despite occasional exceptions. Fourth, we considered exposure a binary notion and, due to the nature of the data, do not have well-defined start-of-exposure. Doses could be considered, perhaps on an ordinal scale, if the interest revolves around dose-related ADRs; the lack of well-defined exposure time could be mitigated if doorstep medication profiles were based on data from the Danish Drug Statistics Register [73] (unavailable to us when conducting this study.) Third, word embeddings are powerful but not magical: the method clearly links clinical terms with similar meanings (even if lexicographically very different) to similar medications profiles, but the embedding model has difficulties

with i.a. rare variations: these yield different embedding vectors resulting in noisy signal profiles fitting poorly with what is expected. However, rarity of terms also hampers other kinds of association-mining or disproportionality-analytic techniques, and our method might even be less prone because few mentions might be enough to at least hint at relevant "clinical cousins" (terms that mean at approximately the same thing.) Fourth, even if the doorstep medication profiles are correct, we have no records of exposure to over-the-counter and herbal drugs and we have to assume patients be compliant, just as any study using secondary data. Finally, we only had data on inpatients who were not, generally, admitted due to side effects. Inpatients are not representative of the general population and so, with the data at our disposal, the safety signals might be somewhat conditional on frailty to some extent.

Conclusion

Combining various flavours of machine learning and data scientific tools we have built an end-to-end pipeline for mining associations between free-text information and medication exposure without the need for manual curation. We achieve this by turning things upside down, predicting not the likely outcome of a range of exposures, but the likely exposures for one or several outcomes of interest.

The congruence analysis suggests that the method pick up pertinent information, even when supplied with synonyms, and with 8% of single-drug and 14% of drug-pair signals being possibly undescribed side effects, it provides a hit rate appropriate for its purpose: shortlisting few relevant signals from thousands of noisy signals [27]. These shortlists would then undergo review by pharmacologists, pharmacists or other pharmacovigilance experts [5,27] to elicit truly unknown side effects or aid substantiating/refuting suspected side effects emerging from e.g. spontaneous case reports.

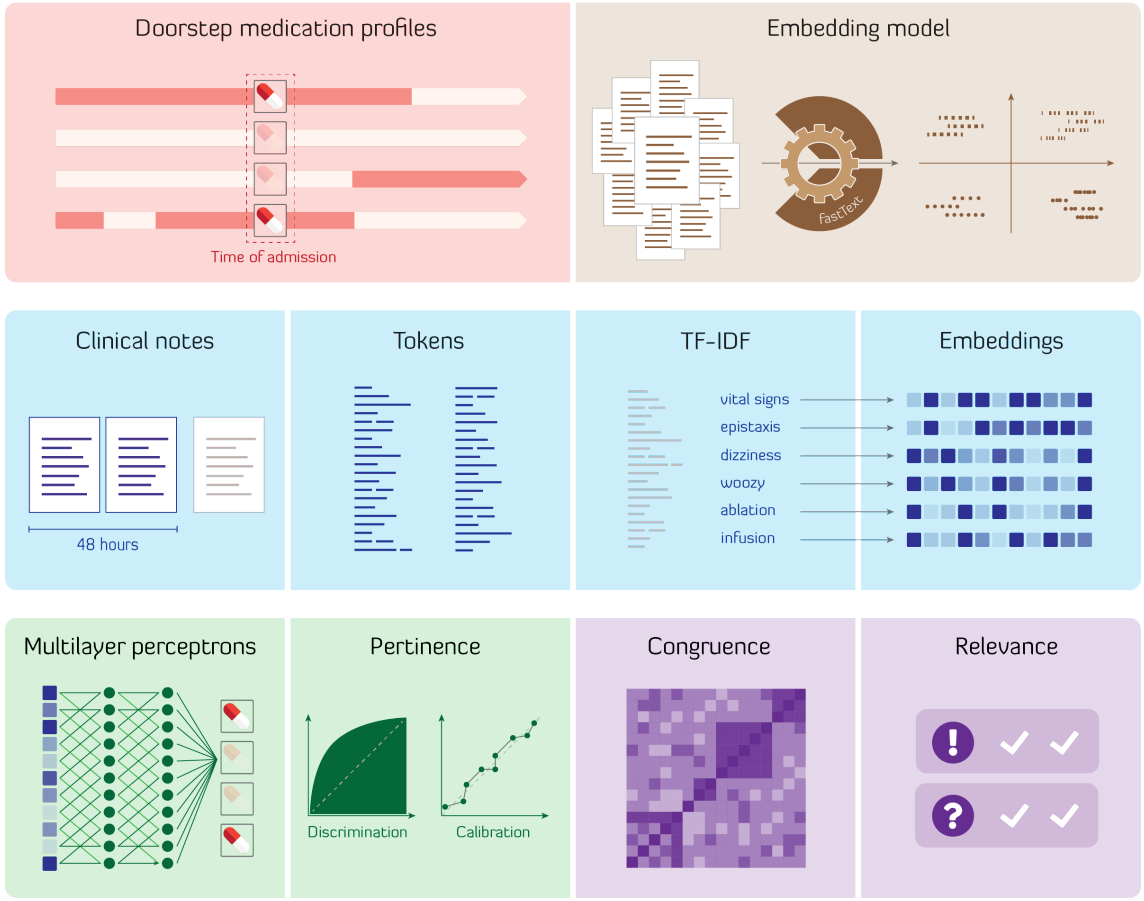
Our approach is original in the field of side effect detection and helps overcome many limitations of NLP methods relying on curated data including being language-agnostic. Crucially, this makes our method appealing in settings that must make sense of non-English free text for pharmacovigilance while lending itself well to alternative use cases, e.g., patient-level decision-making support and drug repurposing.

Acknowledgements

The authors would like to thank DrugBank for granting access to their database. Figure 1 contains various Font Awesome icons (https://fontawesome.com/license).

Figures

Figure 1: Schematic illustration of the end-of-end pipeline, see sections with corresponding headings in main text for details: the blue areas correspond to *Operationalisation of clinical notes*, the green to *Training the signal detection component* and the purple to *Evaluating the safety signals*. The red and blue areas illustrate data capture from a single patient.



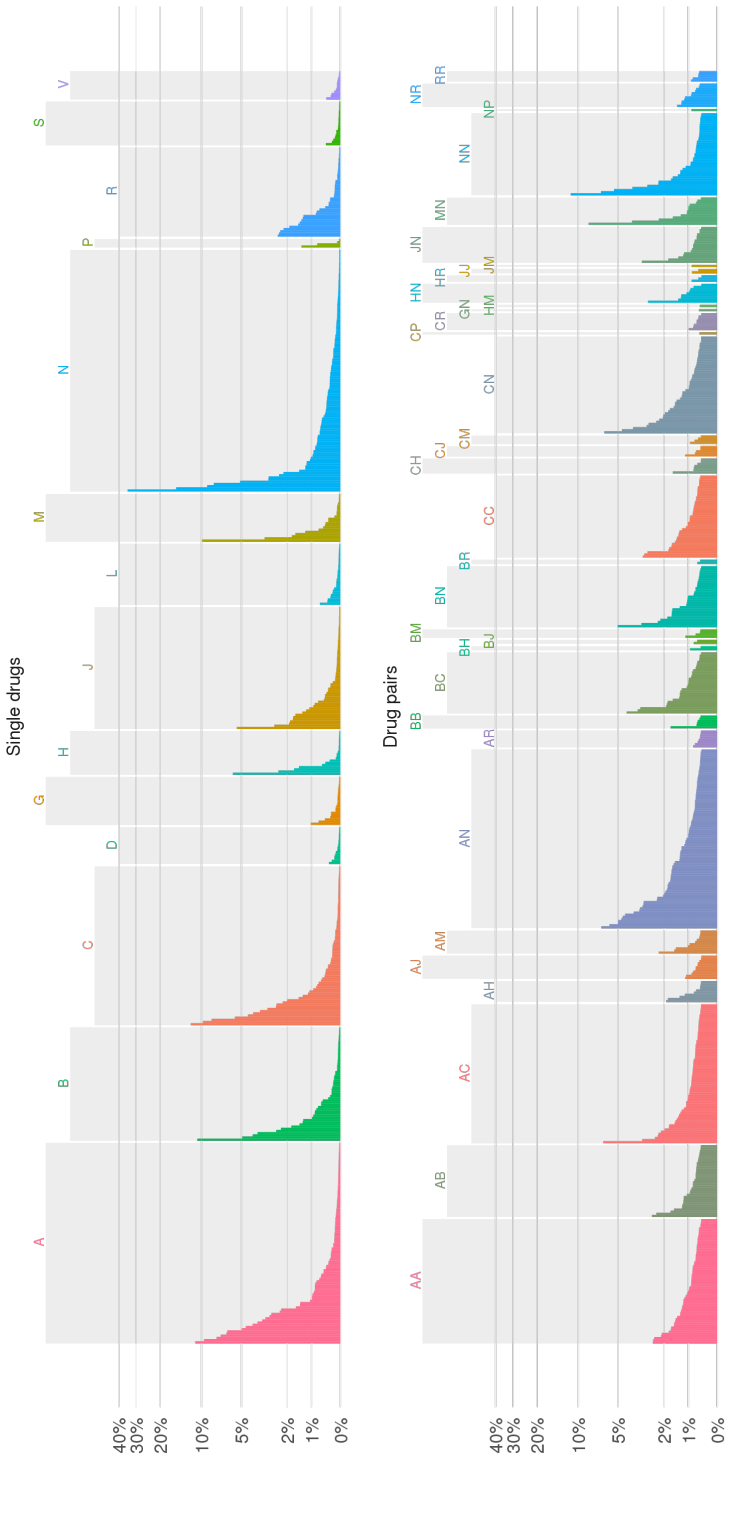


Figure 2: Proportions of included visits with all 571 single drugs (top panel) and the top-571 two-way drug combinations (lower panel), by anatomical drug classes (ATC level 1). Colours in the lower panel have come about by additive mixing of the drug-class colours used in the top panel. The vertical scale is pseudo-log-transformed (linear between 0% and 1%). 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, excluding 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

Figure 3: Fingerprint plots of the 23 main UKU terms and their 571 single-drug signals. Inner circles: each wedge represents one drug and transparency the signal score. Outer circles: colours represent anatomical drug classes (ATC level 1), see legend. See caption of figure 2 for drug-class names.

Figure 4: Mean-adjusted cosine similarities between signal pairs. Rows and columns show pairwise similarities between signal profiles for specific terms. Dark blue squares signify agreement between blocks of terms (red represent disagreement). Black and white margin bars represent UKU side-effect terms, and columns/rows within the span of one bar are synonyms. The cosine similarity of two identical signals equals 1 (e.g. the diagonal).

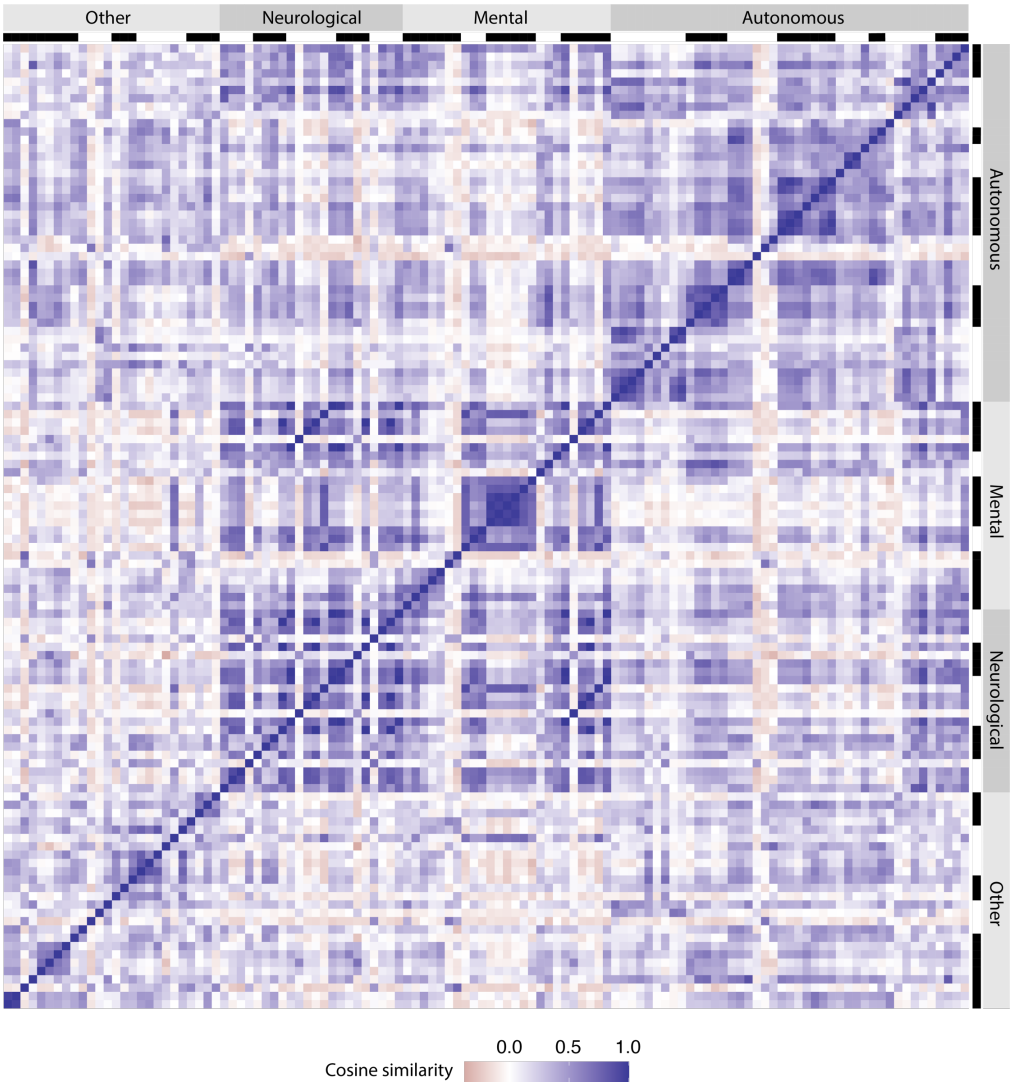
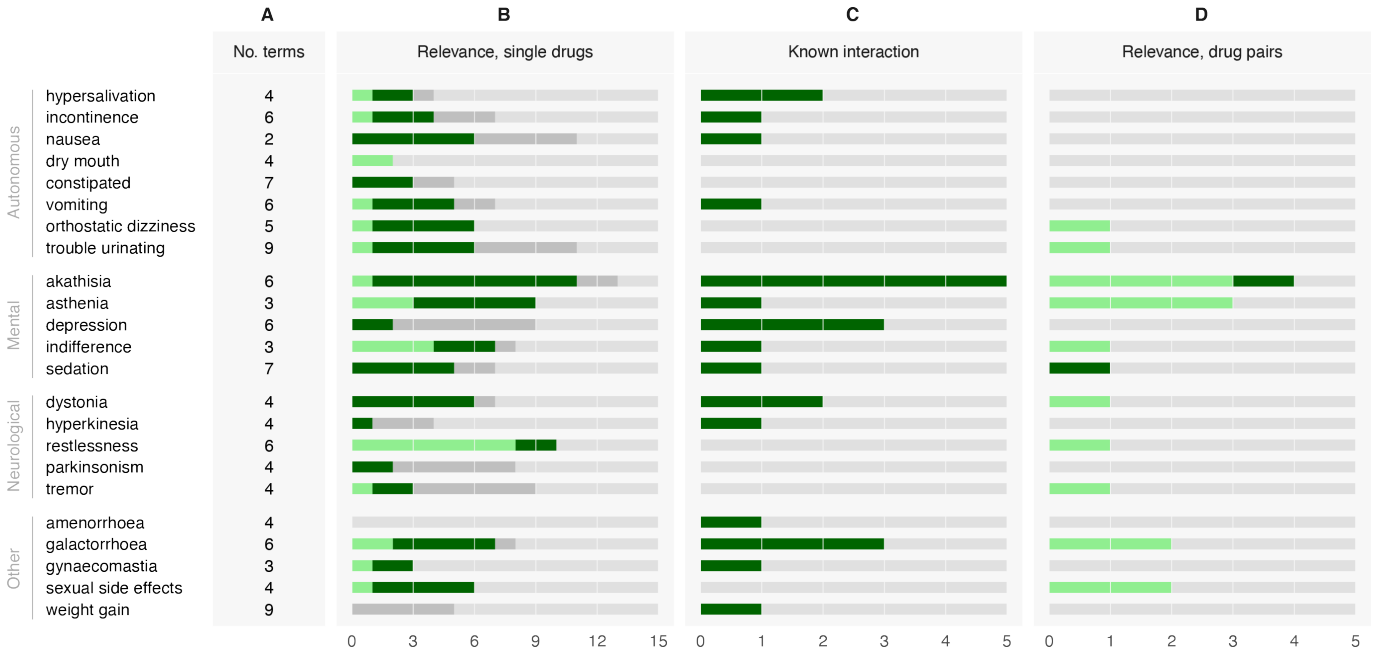


Figure 5: Main UKU terms by domain. Panel A shows the number of terms used in congruence analysis (total = 116). Panel B comprises all 345 single-drug assessments (23 terms x 5 single-drug signals = 115; 23 terms x 5 drug-pair signals x 2 drugs per pair = 230). Light green: reaction possibly caused by single-drug (panel B) or drug-pair (panel D) exposure. Dark green: known reaction (panels B+D) or interaction (panel C). Dark grey: protopathic or indication bias. Light grey: spurious signal. Horisontal scales in panels B-D are counts.



References

1. Delamothe, T. Reporting Adverse Drug Reactions. *BMJ: British Medical Journal* **304,** 465–465 (1992).
2. Nebeker, J. R., Barach, P. & Samore, M. H. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Ann Intern Med* **140,** 795–801 (2004).
3. Human Medicines Evaluation Unit. Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. 1995. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use_en-15.pdf. Accessed 5 October 2021.
4. Edwards, I. R. & Aronson, J. K. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* **356,** 1255–1259 (2000).
5. Lindquist, M. VigiBase, the WHO Global ICSR Database System: Basic Facts. *Drug Information Journal* **42,** 409–419 (2008).
6. Edwards, I. R. An agenda for UK clinical pharmacology: Pharmacovigilance. *Br J Clin Pharmacol* **73,** 979–982 (2012).
7. Alvarez-Requejo, A. et al. Under-reporting of adverse drug reactions. Estimate based on a spontaneous reporting scheme and a sentinel system. *Eur J Clin Pharmacol* **54,** 483–488 (1998).
8. Moride, Y., Haramburu, F., Requejo, A. A. & Bégaud, B. Under-reporting of adverse drug reactions in general practice. *Br J Clin Pharmacol* **43,** 177–181 (1997).
9. Patrignani, A. et al. $[$Under-reporting of adverse drug reactions, a problem that also involves medicines subject to additional monitoring. Preliminary data from a single-center experience on novel oral anticoagulants]. *G Ital Cardiol (Rome)* **19,** 54–61 (2018).
10. Danish Medicines Agency Bivirkningsindberetninger om afhængighed ved tramadol: Gennemgang og analyse af danske indberetninger. Preprint at https://laegemiddelstyrelsen.dk/da/nyheder/2018/ny-rapport-om-bivirkningsindberetninger-om-den-smertestillende-medicin-tramadol (2018).
11. Stergiopoulos, S., Fehrle, M., Caubel, P., Tan, L. & Jebson, L. Adverse Drug Reaction Case Safety Practices in Large Biopharmaceutical Organizations from 2007 to 2017: An Industry Survey. *Pharmaceut Med* **33,** 499–510 (2019).
12. Aggarwal, C. C. *Data Mining – The Textbook* (Springer, , 2015).
13. Goldberg, Y. *Neural Network Methods in Natural Language Processing* (1st ed., {Morgan \& Claypool Publishers}, , 2017).
14. Tatonetti, N. P., Ye, P. P., Daneshjou, R. & Altman, R. B. Data-Driven Prediction of Drug Effects and Interactions. *Science Translational Medicine* **4,** 125ra31–125ra31 (2012).
15. Iyer, S. V., Lependu, P., Harpaz, R., Bauer-Mehren, A. & Shah, N. H. Learning signals of adverse drug-drug interactions from the unstructured text of electronic health records. *AMIA Summits on Translational Science proceedings* **2013,** 83–87 (2013).
16. Iyer, S. V., Harpaz, R., LePendu, P., Bauer-Mehren, A. & Shah, N. H. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association : JAMIA* **21,** 353–362 (2014).
17. Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M. & Ananiadou, S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *J Am Med Inform Assoc* **27,** 39–46 (2020).
18. Eriksson, R., Jensen, P. B., Frankild, S., Jensen, L. J. & Brunak, S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *J Am Med Inform Assoc* **20,** 947–953 (2013).
19. Schmidt, M. et al. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin. Epidemiol.* **7,** 449–490 (2015).
20. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* **5,** 135–146 (2017).
21. Thomas, C. E., Jensen, P. B., Werge, T. & Brunak, S. Negation scope and spelling variation for text-mining of Danish electronic patient records. In: *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)* (2014).
22. Bird, S., Loper, E. & Klein, E. *Natural Language Processing with Python* ({O'Reilly Media Inc.}, , 2009).
23. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge University Press, New York NY, USA, 2008).
24. Steyerberg, E. W. & Harrell, F. E. J. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* **69,** 245–247 (2016).
25. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal* **35,** 1925–1931 (2014).
26. Leon, S. J. *Linear algebra with applications* (7th ed., Pearson Prentice Hall, , 2006).
27. Trifirò, G. et al. Data mining on electronic health record databases for signal detection in pharmacovigilance:

which events to monitor? *Pharmacoepidemiol Drug Saf* **18,** 1176–1184 (2009).
28. Lingjærde, O., Ahlfors, U. G., Bech, P., Dencker, S. J. & Elgen, K. The UKU side effect rating scale: A new comprehensive rating scale for psychotropic drugs and a cross-sectional study of side effects in neuroleptic-treated patients. *Acta Psychiatrica Scandinavica* **76,** 1–100 (1987).
29. Jürgens, G. et al. Effect of Routine Cytochrome P450 2D6 and 2C19 Genotyping on Antipsychotic Drug Persistence in Patients With Schizophrenia: A Randomized Clinical Trial. *JAMA Network Open* **3,** e2027909–e2027909 (2020).
30. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46,** D1074–D1082 (2017).
31. Aagaard, L. & Kristensen, M. B. [The national drug interactions database (in Danish)]. *Ugeskrift for laeger* **167,** 3283–3286 (2005).
32. Faillie, J.-L. Indication bias or protopathic bias? *British Journal of Clinical Pharmacology* **80,** 779–780 (2015).
33. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20,** 37–46 (1960).
34. Benchimol, E. I. et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* **12,** e1001885 (2015).
35. Masnoon, N., Shakib, S., Kalisch-Ellett, L. & Caughey, G. E. What is polypharmacy? A systematic review of definitions. *BMC Geriatry* **17,** 230 (2017).
36. Hult, S. et al. A Feasibility Study of Drug-Drug Interaction Signal Detection in Regular Pharmacovigilance. *Drug Saf* **43,** 775–785 (2020).
37. Warrer, P., Hansen, E. H., Juhl-Jensen, L. & Aagaard, L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br J Clin Pharmacol* **73,** 674–684 (2012).
38. Luo, Y. et al. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Saf* **40,** 1075–1089 (2017).
39. Juhlin, K., Star, K. & Norén, G. N. A method for data-driven exploration to pinpoint key features in medical data and facilitate expert review. *Pharmacoepidemiology and Drug Safety* **26,** 1256–1265 (2017).
40. Norén, G. N., Sundberg, R., Bate, A. & Edwards, I. R. A statistical methodology for drug–drug interaction surveillance. *Statistics in Medicine* **27,** 3057–3070 (2008).
41. Polepalli Ramesh, B. et al. Automatically Recognizing Medication and Adverse Event Information From Food and Drug Administration's Adverse Event Reporting System Narratives. *JMIR Med Inform* **2,** e10 (2014).
42. Maciejewski, M. et al. Reverse translation of adverse event reports paves the way for de-risking preclinical off-targets. *Elife* **6,** (2017).
43. Dewulf, P., Stock, M. & De Baets, B. Cold-Start Problems in Data-Driven Prediction of Drug-Drug Interaction Effects. *Pharmaceuticals (Basel)* **14,** (2021).
44. Masumshah, R., Aghdam, R. & Eslahchi, C. A neural network-based method for polypharmacy side effects prediction. *BMC Bioinformatics* **22,** 385 (2021).
45. Korkontzelos, I. et al. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *J Biomed Inform* **62,** 148–158 (2016).
46. Chen, X. et al. Mining Adverse Drug Reactions in Social Media with Named Entity Recognition and Semantic Methods. *Stud Health Technol Inform* **245,** 322–326 (2017).
47. Rezaallah, B., Lewis, D. J., Pierce, C., Zeilhofer, H.-F. & Berg, B.-I. Social Media Surveillance of Multiple Sclerosis Medications Used During Pregnancy and Breastfeeding: Content Analysis. *J Med Internet Res* **21,** e13003 (2019).
48. Nikfarjam, A. et al. Early Detection of Adverse Drug Reactions in Social Health Networks: A Natural Language Processing Pipeline for Signal Detection. *JMIR Public Health Surveill* **5,** e11264 (2019).
49. Gavrielov-Yusim, N. et al. Comparison of text processing methods in social media-based signal detection. *Pharmacoepidemiol Drug Saf* **28,** 1309–1317 (2019).
50. Alvaro, N. et al. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *J Biomed Inform* **58,** 280–287 (2015).
51. Alvaro, N., Miyao, Y. & Collier, N. TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. *JMIR Public Health Surveill* **3,** e24 (2017).
52. Jiang, K., Chen, T., Calix, R. A. & Bernard, G. R. Prediction of Personal Experience Tweets of Medication Use via Contextual Word Representations(). *Annu Int Conf IEEE Eng Med Biol Soc* **2019,** 6093–6096 (2019).
53. Liu, J., Zhao, S. & Zhang, X. An ensemble method for extracting adverse drug events from social media. *Artif Intell Med* **70,** 62–76 (2016).
54. Emadzadeh, E., Sarker, A., Nikfarjam, A. & Gonzalez, G. Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology. *AMIA Annu Symp Proc* **2017,** 679–688 (2017).
55. Cocos, A., Fiks, A. G. & Masino, A. J. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J Am Med Inform Assoc* **24,** 813–821 (2017).

56. Bian, J., Topaloglu, U. & Yu, F. Towards Large-scale Twitter Mining for Drug-related Adverse Events. In: *SHB '12: Proceedings of the 2012 international workshop on Smart health and wellbeing* (2012).
57. Abdellaoui, R., Schück, S., Texier, N. & Burgun, A. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help? *JMIR Public Health Surveill* **3,** e36 (2017).
58. Carbonell, P., Mayer, M. A. & Bravo, À. Exploring brand-name drug mentions on Twitter for pharmacovigilance. *Stud Health Technol Inform* **210,** 55–59 (2015).
59. Gattepaille, L. M. How far can we go with just out-of-the-box BERT models? In: *Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task* (2020).
60. Eshleman, R. & Singh, R. Leveraging graph topology and semantic context for pharmacovigilance through twitter-streams. *BMC Bioinformatics* **17,** 335 (2016).
61. Eriksson, R., Werge, T., Jensen, L. J. & Brunak, S. Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population. *Drug Saf* **37,** 237–247 (2014).
62. Sørup FKH. Exploring Associations Between Text Mined Adverse Events and Antipsychotic Drug use (PhD thesis). 2019.
63. Oronoz, M., Gojenola, K., Perez, A., de Ilarraza, A. D. & Casillas, A. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *J Biomed Inform* **56,** 318–332 (2015).
64. Segura-Bedmar, I. & Martinez, P. Pharmacovigilance through the development of text mining and natural language processing techniques. *J Biomed Inform* **58,** 288–291 (2015).
65. Matsuda, S. et al. Analysis of Patient Narratives in Disease Blogs on the Internet: An Exploratory Study of Social Pharmacovigilance. *JMIR Public Health Surveill* **3,** e10 (2017).
66. Ujiie, S., Yada, S., Wakamiya, S. & Aramaki, E. Identification of Adverse Drug Event-Related Japanese Articles: Natural Language Processing Analysis. *JMIR Med Inform* **8,** e22661 (2020).
67. Usui, M. et al. Extraction and Standardization of Patient Complaints from Electronic Medication Histories for Pharmacovigilance: Natural Language Processing Analysis in Japanese. *JMIR Med Inform* **6,** e11021 (2018).
68. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* **6,** 60 (2019).
69. Workman, T. E., Divita, G., Shao, Y. & Zeng-Treitler, Q. A Proficient Spelling Analysis Method Applied to a Pharmacovigilance Task. *Stud Health Technol Inform* **264,** 452–456 (2019).
70. Lavertu, A. & Altman, R. B. RedMed: Extending drug lexicons for social media applications. *J Biomed Inform* **99,** 103307 (2019).
71. Chollet, F. *Deep Learning with Python* (Manning Publications Co., New York, USA, 2018).
72. Kessing, L. V. et al. Antihypertensive Drugs and Risk of Depression. *Hypertension* **76,** 1263–1279 (2020).
73. Gregersen, R., Wiingreen, R. & Rosenberg, J. [Health-related register-based research in Denmark (in Danish)]. *Ugeskr Laeger* **180,** (2018).

Eliciting side effects from clinical notes: language-agnostic pharmacovigilant text mining

SUPPLEMENT

Benjamin Skov Kaas-Hansen MD, MSc — Davide Placido MSc
Cristina Leal Rodríguez MSc, PhD — Hans-Christian Thorsen-Meyer MD, PhD
Simona Gentile MD — Anna Pors Nielsen MD
Søren Brunak MSc, PhD — Gesche Jürgens MD, PhD — Stig Ejdrup Andersen MD, PhD

Figure S1: Intercepts (x axis) and slopes (y axis) of linear regressions of the calibration curves in the internal validation sets. Colour represents AUROC (0.5 corresponds to random guessing, 1.0 to perfect discrimination). Models with intercept > 0 tend to have slopes < 1 and vice-versa, as a compensatory mechanism. Models represented by points inside the rectangle yield pertinent signals.

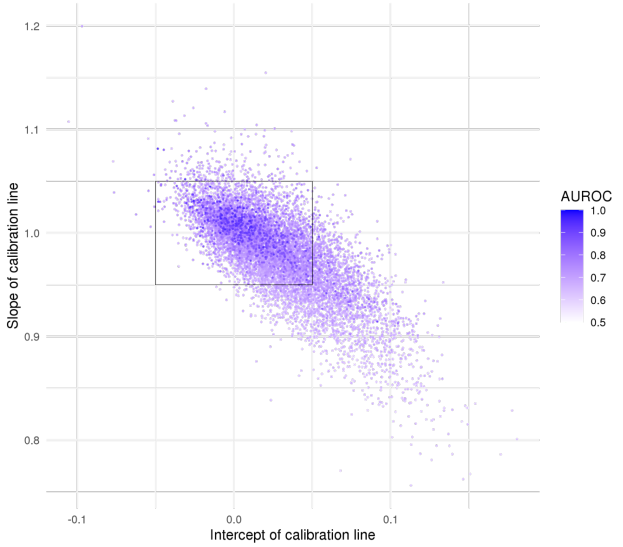
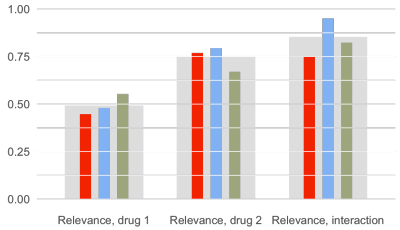


Figure S2: Cohen's kappa for each rater pair (coloured bars) and overall (shaded, wide bar in the background), by item.



[ ]

Acronyms

ADE adverse drug event. 23

ADR adverse drug reaction. 9, 23–25, 27, 33, 62, 73, 76, 80

ATC Anatomic Therapeutic Chemical classification. 33, 34, 58

AUROC area under the ROC curve. 50, 59

BCC B-Data Clinical Chemistry Laboratory System. 34, 69

CDM common data model. 67, 75, 78–80

CPU central processing unit. 72

CRN civil registration number. 31

CRS the National Civil Registration System. 31, 32

CYP cytochrom P450. 57

DDI drug-drug interaction. 11, 27, 55–57, 73

DID Danish Drug Interactions Database. 55

DMA the Danish Medicines Agency. 25, 55

eGFR estimated glomerular filtration rate. 34, 58

EHR electronic health record. 26, 63, 67, 77

EMA the European Medicines Agency. 25, 80

EMR electronic medical record. 10, 26, 35, 63, 76

EPM Electronic Patient Medication. 68, 79

EPM1 Electronic Patient Medication 1. 33, 67

EPM3 Electronic Patient Medication 3. 33, 67

EPR electronic patient record. 26

ETL extract-transform-load. 33, 67, 68, 70, 72, 80

EU European Union. 25

FAERS the FDA Adverse Event Reporting System. 25

GAM generalised additive model. 40

GLM generalised linear model. 40–44, 48, 66

HMM Hidden Markov Model. 64

ICD10 the 10th revision of the International Classification of Disease. 37, 70
ICSR individual case safety report. 24–26, 62, 73, 75, 81
ICU intensive care unit. 32, 33, 67
IQR inter-quartile range. 56
IUPAC International Union of Pure and Applied Chemistry. 71

LABKA Clinical Laboratory Information System (in Danish: sygehus-Laboratorier, Klinisk Biokemiske Afdelinger). 34

MAR missing-at-random. 79
MCAR missing-completely-at-random. 79
MCC Matthew's correlation coefficient. 50
MedDRA the Medical Dictionary for Regulatory Activities. 37
MLP multilayer perceptron. 41–43, 46–48, 52, 57–61, 64, 73
MNAR missing-not-at-random. 79

NCSP NOMESCO Classification of Surgical Procedures. 70
NLP natural language processing. 27, 35, 37, 59, 62, 80
NPR the Danish National Patient Register. 32, 58, 69, 70
NPU Nomenclature, Properties and Units. 69, 71
NSAID non-steroidal anti-inflammatory drug. 57

OHDSI Observational Health Data Sciences and Informatics program. 66, 67, 72, 78, 80, 81
OMOP Observational Medical Outcomes Partnership. 67, 75, 79, 80

RCT randomised controlled trial. 27, 28, 40, 65–67
ROC receiver operating characteristic. 50

SHAP SHapley Additive exPlanation. 28, 52, 58, 59, 65, 66
SKS Sygehusvæsenets Klassifikationssystem. 32, 70
SLOR shrinkage log odds ratio. 63, 73
SPD standardised difference in proportions. 45, 55, 63
SRS spontaneous reporting system. 25, 75

UK United Kingdom. 24
UKU Udvalg for KLiniske Undersøgelser. 61
UMC Uppsala Monitoring Centre. 24, 34
USA United States of America. 25, 34

WHO World Health Organization. 26, 33

xAI explainable artificial intelligence. 28, 52

References

1. World Health Organization. *The safety of medicines in public health programmes: pharmacovigilance an essential tool* tech. rep. (WHO Collaborating Centre for International Drug Monitoring, 2006). https://digicollections.net/medicinedocs/documents/s14085e/s14085e.pdf.
2. Kim, C. & Prasad, V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. *JAMA Internal Medicine* **175,** 1992–1994. doi:10.1001/jamainternmed.2015.5868 (Dec. 2015).
3. Davis, C. *et al.* Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13. eng. *BMJ* **359,** j4530. doi:10.1136/bmj.j4530 (Oct. 2017).
4. Kornholt, J. & Christensen, M. B. Prevalence of polypharmacy in Denmark. eng. *Danish Medical Journal* **67** (June 2020).
5. Onakpoya, I. J., Heneghan, C. J. & Aronson, J. K. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Medicine* **14,** 10. doi:10.1186/s12916-016-0553-2 (2016).
6. Wood, S. *Generalized Additive Models: An Introduction with R* 2nd ed. (Chapman and Hall/CRC, 2017).
7. Rodríguez, C. L. *Mining Polypharmacy Data from Electronic Health Records: A data driven approach for the investigation of polypharmacy, drug interactions and drug dosage* PhD thesis (University of Copenhagen, 2021).
8. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine* **384,** 693–704. doi:10.1056/NEJMoa2021436 (2021).
9. The REMAP-CAP Investigators. Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *New England Journal of Medicine* **384.** PMID: 33631065, 1491–1502. doi:10.1056/NEJMoa2100433 (2021).
10. Edwards, I. R. A New Erice Report Considering the Safety of Medicines in the 21st Century. *Drug Safety* **40,** 845–849. doi:10.1007/s40264-017-0571-9 (2017).
11. Kaas-Hansen, B. S. *et al. An OMOP-based tool for surveying and visualising concurrent drug exposure and renal function* in *The Second European OHDSI Symposium 2019 Poster Session* (2019). https://www.ohdsi-europe.org/index.php/symposium/16-archive-symposium-2019.
12. Nebeker, J. R., Barach, P. & Samore, M. H. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. eng. *Annals of Internal Medicine* **140,** 795–801. doi:10.7326/0003-4819-140-10-200405180-00009 (2004).
13. Human Medicines Evaluation Unit. *Clinical Safety Data Management: Definitions and Standards for Expedited Reporting* tech. rep. (European Agency for the Evaluation of Medicinal Products, 1995). https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use_en-15.pdf.
14. Edwards, I. R. & Aronson, J. K. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* **356,** 1255–1259. doi:10.1016/S0140-6736(00)02799-9 (2000).
15. *The Drug Development Process* [Online; accessed 28. Oct. 2021]. 2018. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process.
16. Edwards, I. R. An agenda for UK clinical pharmacology: Pharmacovigilance. *British Journal of Clinical Pharmacology* **73,** 979–982. doi:10.1111/j.1365-2125.2012.04249.x (2012).
17. Abraham, J. & Lewis, G. *Regulating medicine in Europe: competition, expertise and public health* (Routledge, 2000).

18. World Health Organization. *Regulation and prequalification* https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance.
19. World Health Organization. *Pharmacovigilance: ensuring the safe use of mecicines* tech. rep. (World Health Organization, 2004).
20. Pontes, H., Clément, M. & Rollason, V. Safety signal detection: the relevance of literature review. eng. *Drug Safety* **37,** 471–479. doi:10.1007/s40264-014-0180-9 (2014).
21. Hult, S. *et al.* A Feasibility Study of Drug-Drug Interaction Signal Detection in Regular Pharmacovigilance. eng. *Drug Safety* **43,** 775–785. doi:10.1007/s40264-020-00939-y (June 2020).
22. Candore, G. *et al.* Comparison of statistical signal detection methods within and across spontaneous reporting databases. eng. *Drug Safety* **38,** 577–587. doi:10.1007/s40264-015-0289-5 (June 2015).
23. U.S. Food and Drug Administration. *Questions and Answers on FDA's Adverse Event Reporting System (FAERS)* [Online; accessed 6. Oct. 2021]. 2018. https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers.
24. Grandvuillemin, A. *et al.* French Pharmacovigilance Public System and COVID-19 Pandemic. *Drug Safety.* doi:10.1007/s40264-020-01034-y (2020).
25. Greinacher, A. *et al.* Thrombotic Thrombocytopenia after ChAdOx1 nCov-19 Vaccination. *New England Journal of Medicine* **384,** 2092–2101. doi:10.1056/NEJMoa2104840 (2021).
26. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46,** D1074–D1082. doi:10.1093/nar/gkx1037 (Nov. 2017).
27. *Rene agonister - information til sundhedsfaglige - Medicin.dk* [Online; accessed 11. Oct. 2021]. 2021. https://pro.medicin.dk/Laegemiddelgrupper/Grupper/227010#a300.
28. Danish Medicines Agency. *Medicines with stricter reporting requirements for doctors, dentists, veterinarians, midwives and prescribing pharmacists* Online. [Online; accessed 6. Oct. 2021]. 2021. https://laegemiddelstyrelsen.dk/en/sideeffects/side-effects-of-medicines/medicines-with-stricter-reporting-requirements.
29. *Lægemiddelstyrelsens årsrapport for overvågning af bivirkninger 2016 [in Danish]* ISBN: 978-87-92390-19-6 (Danish Medicines Agency, 2017).
30. Hazell, L. & Shakir, S. A. W. Under-reporting of adverse drug reactions: a systematic review. eng. *Drug Safety* **29,** 385–396 (2006).
31. Lander, A. R. *et al.* Introducing an adverse drug event manager. *European Journal of Hospital Pharmacy: Science and Practice* **20,** 78–81. doi:10.1136/ejhpharm-2012-000171 (2013).
32. Sørup, F. K. H., Jacobsen, C. B. & Jimenez-Solem, E. Increasing the Number of Spontaneous ADE Reports in a Danish Region: A Retrospective Analysis. *Pharmaceutical Medicine* **29,** 211–217. doi:10.1007/s40290-015-0102-x (2015).
33. Vinther, S. *et al.* An adverse drug event manager facilitates spontaneous reporting of adverse drug reactions. *Danish Medical Journal* **64** (2017).
34. Meyboom, R. H. *et al.* Principles of signal detection in pharmacovigilance. eng. *Drug Saf* **16,** 355–365. doi:10.2165/00002018-199716060-00002 (June 1997).
35. Hauben, M. & Aronson, J. K. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. eng. *Drug Safety* **32,** 99–110. doi:10.2165/00002018-200932020-00003 (2009).
36. Council for International Organizations of Medical Sciences. *Practical Aspects of Signal Detection in Pharmacovigilance: Report of CIOMS Working Group VIII* tech. rep. [Online; accessed 2. Nov. 2021] (World Health Organization, 2009). https://cioms.ch/wp-content/uploads/2018/03/WG8-Signal-Detection.pdf.
37. Delamothe, T. Reporting Adverse Drug Reactions. eng. *British Medical Journal* **304,** 465–465. https://www.jstor.org/stable/29714599 (1992).
38. Harpaz, R *et al.* Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther* **91,** 1010–1021. doi:10.1038/clpt.2012.50 (June 2012).
39. Korkontzelos, I. *et al.* Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. eng. *Journal of Biomedical Informatics* **62,** 148–158. doi:10.1016/j.jbi.2016.06.007 (2016).

40. Chen, X. *et al.* Mining Adverse Drug Reactions in Social Media with Named Entity Recognition and Semantic Methods. eng. *Studies in health technology and informatics* **245,** 322–326. http://ebooks.iospress.nl/Extern/EnterMedLine.aspx?ISSN=0926-9630&Volume=245&SPage=322 (2017).

41. Rezaallah, B., Lewis, D. J., Pierce, C., Zeilhofer, H.-F. & Berg, B.-I. Social Media Surveillance of Multiple Sclerosis Medications Used During Pregnancy and Breastfeeding: Content Analysis. *Journal of Medical Internet Research* **21,** e13003. doi:10.2196/13003 (2019).

42. Nikfarjam, A. *et al.* Early Detection of Adverse Drug Reactions in Social Health Networks: A Natural Language Processing Pipeline for Signal Detection. eng. *JMIR Public Health Surveillance* **5,** e11264. doi:10.2196/11264 (2019).

43. Gavrielov-Yusim, N. *et al.* Comparison of text processing methods in social media-based signal detection. eng. *Pharmacoepidemiology and Drug Safety* **28,** 1309–1317. doi:10.1002/pds.4857 (Oct. 2019).

44. Alvaro, N. *et al.* Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *Journal of Biomedical Informatics* **58,** 280–287. doi:10.1016/j.jbi.2015.11.004 (2015).

45. Alvaro, N., Miyao, Y. & Collier, N. TwiMed: Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms, and Their Relations. eng. *JMIR Public Health Surveillance* **3,** e24. doi:10.2196/publichealth.6396 (May 2017).

46. Jiang, K., Chen, T., Calix, R. A. & Bernard, G. R. Prediction of Personal Experience Tweets of Medication Use via Contextual Word Representations. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2019,** 6093–6096. doi:10.1109/EMBC.2019.8856753 (July 2019).

47. Liu, J., Zhao, S. & Zhang, X. An ensemble method for extracting adverse drug events from social media. *Artificial intelligence in medicine* **70,** 62–76. doi:10.1016/j.artmed.2016.05.004 (June 2016).

48. Emadzadeh, E., Sarker, A., Nikfarjam, A. & Gonzalez, G. Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology. *AMIA Annual Symposium proceedings* **2017,** 679–688. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977584/ (2017).

49. Cocos, A., Fiks, A. G. & Masino, A. J. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association* **24,** 813–821. doi:10.1093/jamia/ocw180 (July 2017).

50. Bian, J., Topaloglu, U. & Yu, F. *Towards Large-scale Twitter Mining for Drug-related Adverse Events.* in *SHB '12: Proceedings of the 2012 international workshop on Smart health and wellbeing* **2012** (2012), 25–32. doi:10.1145/2389707.2389713.

51. Abdellaoui, R., Schück, S., Texier, N. & Burgun, A. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help? *JMIR Public Health and Surveillance* **3,** e36. doi:10.2196/publichealth.6577 (June 2017).

52. Carbonell, P., Mayer, M. A. & Bravo, À. Exploring brand-name drug mentions on Twitter for pharmacovigilance. *Studies in health technology and informatics* **210,** 55–59. doi:10.3233/978-1-61499-512-8-55 (2015).

53. Norén, G. N., Hopstadius, J., Bate, A., Star, K. & Edwards, I. R. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery* **20,** 361–387. doi:10.1007/s10618-009-0152-3 (2010).

54. Wilson, A. M., Thabane, L. & Holbrook, A. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology* **57,** 127–134. doi:10.1046/j.1365-2125.2003.01968.x (2004).

55. Star, K., Watson, S., Sandberg, L., Johansson, J. & Edwards, I. R. Longitudinal medical records as a complement to routine drug safety signal analysis. *Pharmacoepidemiology and Drug Safety* **24,** 486–494. doi:10.1002/pds.3739 (May 2015).

56. Liu, M. *et al.* Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association* **20,** 420–426. doi:10.1136/amiajnl-2012-001119 (May 2013).

57. Coloma, P. M., Trifiro, G., Patadia, V. & Sturkenboom, M. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? *Drug Saf* **36,** 183–197. doi:10.1007/s40264-013-0018-x (Mar. 2013).

58. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13,** 395–405. doi:10.1038/nrg3208 (2012).

59. Kierkegaard, P. Electronic health record: Wiring Europe's healthcare. *Computer Law & Security Review* **27,** 503–515. doi:10.1016/j.clsr.2011.07.013 (2011).

60. Habib, J. L. EHRs, Meaningful Use, and a Model EMR. *Drug Benefit Trends* **22.** http://web.archive.org/web/20100728133252/http://dbt.consultantlive.com:80/display/article/1145628/1581538 (2010).

61. Garets, D. & Davis, M. *Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference* tech. rep. (HIMSS Analytics, LLC, 2006). http://web.archive.org/web/20110726151554/http://www.himssanalytics.org/docs/WP_EMR_EHR.pdf.

62. Neal, M. J. *Medical Pharmacology at a Glance* ISBN: 9781118902417 (John Wiley & Sons inc., 2015).

63. Iyer, S. V., Harpaz, R., LePendu, P., Bauer-Mehren, A. & Shah, N. H. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association* **21,** 353–362. doi:10.1136/amiajnl-2013-001612 (Apr. 2014).

64. Rothman, K. J., Lash, T. L. & Greenland, S. *Modern Epidemiology* 3rd ed. ISBN: 9781451190052 (Lippincott Williams & Wilkins, 2012).

65. Arnold, K. F. *et al.* Reflections on modern methods: generalized linear models for prognosis and intervention-theory, practice and implications for machine learning. *International Journal of Epidemiology.* doi:10.1093/ije/dyaa049 (May 2020).

66. Wilkinson, J. *et al.* Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health.* doi:10.1016/S2589-7500(20)30200-4 (2020).

67. Porta, M. *A Dictionary of Epidemiology* 5th ed. (eds Greenland, S. & Last, J. M.) (Oxford University Press, 2008).

68. Sørensen, H. T. Editorial - Clinical Epidemiology. *Clinical Epidemiology* **1,** 17–18. doi:10.2147/CLEP.S5309 (2009).

69. Goldberg, Y. *Neural Network Methods in Natural Language Processing* 1st. ISBN: 978-1627052986 (Morgan & Claypool Publishers, 2017).

70. Chollet, F. *Deep Learning with Python* ISBN: 978-1617294433 (Manning Publications Co., New York, USA, 2018).

71. Ryan, M. *Deep Learning with Structured Data* ISBN: 9781617296727. https://www.manning.com/books/deep-learning-with-structured-data (Manning Publications Co., 2020).

72. Shmueli, G. To Explain or to Predict? *Statistical Science* **25,** 289–310. doi:10.1214/10-STS330 (2010).

73. Lundberg, S. M. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* in *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017), 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

74. Wasserman, L. A. *All of Statistics: A Concise Course in Statistical Inference* 2nd ed. ISBN: 978-0387402727 (Springer, 2014).

75. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (MIT Press, Cambridge (MA), USA, 2016).

76. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* **3,** 210–229. doi:10.1147/rd.33.0210 (July 1959).

77. Aggarwal, C. C. *Data Mining – The Textbook* ISBN: 978-3-319-14141-1. doi:10.1007/978-3-319-14142-8 (Springer, 2015).

78. Russell, S. *Human Compatible: AI and the Problem of Control* ISBN: 978-0-241-33520-8 (Penguin Books Ltd, 2019).

79. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* ISBN: 9780553418811 (Crown, 2016).

80. McLennan, S., Lee, M. M., Fiske, A. & Celi, L. A. AI Ethics Is Not a Panacea. *The American Journal of Bioethics* **20,** 20–22. doi:10.1080/15265161.2020.1819470 (2020).

81. Campello, R. J. G. B., Moulavi, D. & Sander, J. *Density-Based Clustering Based on Hierarchical Density Estimates* in *Advances in Knowledge Discovery and Data Mining* (eds Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), 160–172. ISBN: 978-3-642-37456-2.

82. Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach* ISBN: 9780134610993 (Pearson, Apr. 2021).

83. Schmidt, M. *et al.* The Danish National Patient Registry: a review of content, data quality, and research potential. *Clinical Epidemiology* **7,** 449–490. doi:10.2147/CLEP.S91125 (Nov. 2015).

84. Laugesen, K. *et al.* Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. *Clinical Epidemiology* **13,** 533–554. doi:10.2147/CLEP.S314959 (2021).

85. *Dokumentation* [Online; accessed 7. Oct. 2021]. 2021. https://www.esundhed.dk/Dokumentation.

86. Statistics Denmark. *FOLK1A: Population at the first day of the quarter by region, sex, age and marital status* Online. [Online; accessed 7. Oct. 2021]. www.statbank.dk/FOLK1A.

87. Helweg-Larsen, K. The Danish Register of Causes of Death. *Scandinavian Journal of Public Health* **39,** 26–29. doi:10.1177/1403494811399958 (2011).

88. Ylijoki-Sørensen, S. *Autopsy in unresolved death and accuracy of mortality statistics: analysis of cause of death investigation in Finland and in Denmark* PhD thesis (Aarhus University, Aarhus, 2015).

89. Jensen, T. B. *et al.* Content and validation of the Electronic Patient Medication module (EPM)—the administrative in-hospital drug use database in the Capital Region of Denmark. *Scandinavian Journal of Public Health.* doi:10.1177/1403494818760050 (2018).

90. WHO Collaborating Centre for Drug Statistics Methodology. *WHOCC - Home* [Online; accessed 9. Oct. 2021]. 2021. https://www.whocc.no.

91. Lagerlund, O., Strese, S., Fladvad, M. & Lindquist, M. WHODrug: A Global, Validated and Updated Dictionary for Medicinal Information. *Therapeutic Innovation & Regulatory Science* **54,** 1116–1122. doi:10.1007/s43441-020-00130-6 (2020).

92. U.S. National Library of Medicine. *RxNorm Technical Documentation - Table of Contents* Online. [Online; accessed 9. Oct. 2021]. 2020. https://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.

93. *WHOCC - Structure and principles* Online. [Online; accessed 29. Oct. 2021]. Feb. 2018. https://www.whocc.no/atc/structure_and_principles/.

94. Muse, V. P., Aguayo-Orozco, A. & Brunak, S. *Population-wide analysis of hospital biochemical tests to assess seasonal variation and the need to apply temporal correction of reference limits* In preparation. 2021.

95. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* **150,** 604–612. doi:10.7326/0003-4819-150-9-200905050-00006 (May 2009).

96. Corsonello, A. *et al.* Estimating renal function to reduce the risk of adverse drug reactions. *Drug Safety* **35 Suppl 1,** 47–54. doi:10.1007/BF03319102 (2012).

97. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* ISBN: 978-0-521-86571-5 (Cambridge University Press, New York NY, USA, 2008).

98. Leon, S. J. *Linear algebra with applications* 7th. ISBN: 0-13-185785-1 (Pearson Prentice Hall, 2006).

99. Guo, C. & Berkhahn, F. Entity Embeddings of Categorical Variables. *arXiv (unpublished)* (2016).

100. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. eng. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **78,** 947–1012 (2016).

101. Gong, M. *et al.* Causal Generative Domain Adaptation Networks. *arXiv (unpublished).* https://arxiv.org/abs/1804.04333 (2018).

102. Sani, N., Malinsky, D. & Shpitser, I. Explaining the Behavior of Black-Box Prediction Algorithms with Causal Learning. *arXiv (unpublished).* https://arxiv.org/abs/2006.02482 (2021).

103. Chernozhukov, V., Wuthrich, K. & Zhu, Y. A t-test for synthetic controls. *arXiv (unpublished).* https://arxiv.org/abs/1812.10820 (2021).

104. Efron, B. & Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* ISBN: 978-1-107-14989-2 (Cambridge University Press, London, United Kingdom, 2016).

105. Beniaguev, D., Segev, I. & London, M. Single cortical neurons as deep artificial neural networks. *Neuron* **109,** 2727–2739.e3. doi:doi.org/10.1016/j.neuron.2021.07.002 (2021).

106. Thorsen-Meyer, H.-C. *Prognostication in intensive care using machine learning* English. PhD thesis (University of Copenhagen, 2020).

107. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* **49,** 1225–1231. doi:10.1016/s0895-4356(96)00002-9 (1996).

108. Cox, D. R. Two further applications of a model for binary regression. *Biometrika* **45,** 562–565. doi:10.1093/biomet/45.3-4.562 (Dec. 1958).

109. Agresti, A. *An Introduction to Categorical Data Analysis* ISBN: 0-471-11338-7 (Wiley, 1996).

110. Williamson, T., Eliasziw, M. & Fick, G. H. Log-binomial models: exploring failed convergence. *Emerging themes in epidemiology* **10,** 14. doi:10.1186/1742-7622-10-14 (Dec. 2013).

111. Austin, P. C. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics - Simulation and Computation* **38,** 1228–1234. doi:10.1080/03610910902859574 (2009).

112. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7,** 1–26. doi:10.1214/aos/1176344552 (1979).

113. Mullahy, J. Specification and testing of some modified count data models. *Journal of Econometrics* **33,** 341–365. doi:https://doi.org/10.1016/0304-4076(86)90002-3 (1986).

114. Lambert, D. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* **34,** 1–14. doi:10.1080/00401706.1992.10485228 (1992).

115. Altman, D. G. *Practical Statistics for Medical Research* ISBN: 0-412-27630-5 (Chapman & Hall/CRC, 1999).

116. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B, Methodological* **34,** 187–220 (1972).

117. Steyerberg, E. W. & Harrell, F. E. J. Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology* **69,** 245–247. doi:10.1016/j.jclinepi.2015.04.005 (Jan. 2016).

118. Steyerberg, E. W. *Clinical prediction models: a practical approach to development, validation, and updating* ISBN: 9780387772431 (Springer, New York, 2009).

119. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd. ISBN: 0-387-84857-6 (Springer, New York, 2009).

120. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv (unpublished).* http://arxiv.org/abs/1907.10902 (2019).

121. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **104,** 148–175. doi:10.1109/JPROC.2015.2494218 (Jan. 2016).

122. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* **2,** e489–e492. doi:10.1016/S2589-7500(20)30186-2 (2020).

123. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal* **35,** 1925–1931. doi:10.1093/eurheartj/ehu207 (June 2014).

124. Royston, P. & Parmar, M. K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology* **13,** 152. doi:10.1186/1471-2288-13-152 (Dec. 2013).

125. Baldi, P, Brunak, S, Chauvin, Y, Andersen, C. A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16,** 412–424. doi:10.1093/bioinformatics/16.5.412 (May 2000).

126. Thorsen-Meyer, H.-C. *et al.* Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health.* doi:10.1016/S2589-7500(20)30018-2 (2020).

127. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Medical decision making : an international journal of the Society for Medical Decision Making* **26,** 565–574. doi:10.1177/0272989X06295361 (2006).

128. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352.** doi:10.1136/bmj.i6 (2016).

129. Diprose, W. K. *et al.* Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* **27,** 592–600. doi:10.1093/jamia/ocz229 (Apr. 2020).

130. Zhang, Y. *et al.* An explainable supervised machine learning predictor of acute kidney injury after adult deceased donor liver transplantation. *Journal of Translational Medicine* **19,** 321. doi:10.1186/s12967-021-02990-4 (2021).

131. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* **2,** 749–760. doi:10.1038/s41551-018-0304-0 (2018).

132. Gordon, L., Grantcharov, T. & Rudzicz, F. Explainable Artificial Intelligence for Safe Intraoperative Decision Support. *JAMA Surgery* **154,** 1064–1065. doi:10.1001/jamasurg.2019.2821 (Nov. 2019).

133. Aagaard, L. & Kristensen, M. B. The national drug interactions database [in Danish]. *Ugeskrift for laeger* **167,** 3283–3286. http://europepmc.org/abstract/MED/16138969 (2005).

134. Sloane, D. & Morgan, S. P. An Introduction to Categorical Data Analysis. *Annual Review of Sociology* **22,** 351–375. doi:10.1146/annurev.soc.22.1.351 (1996).

135. Walker, A. *et al.* A tool for assessing the feasibility of comparative effectiveness research. *Comparative effectiveness research (Auckland)* **3,** 11–20. doi:10.2147/CER.S40357 (2013).

136. Stürmer, T. *et al.* A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* **59,** 437–447. doi:10.1016/j.jclinepi.2005.07.004 (May 2006).

137. Schneeweiss, S. *et al.* High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20,** 512–522. doi:10.1097/EDE.0b013e3181a663cc (July 2009).

138. Raghunathan, K., Layton, J. B., Ohnuma, T. & Shaw, A. D. Observational Research Using Propensity Scores. *Advances in Chronic Kidney Disease* **23.** Evidence-Based Medicine, 367–372. doi:10.1053/j.ackd.2016.11.010 (2016).

139. Andersen, S. E. *Health Technology Assessment Perspective on Prescription Writing* PhD thesis (University of Copenhagen, 2002).

140. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* **5,** 135–146. https://transacl.org/ojs/index.php/tacl/article/view/999 (2017).

141. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20,** 37–46. doi:10.1177/001316446002000104 (1960).

142. Trifirò, G. *et al.* EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. *Studies in Health Technology and Informatics* **166,** 25–30. doi:10.3233/978-1-60750-740-6-25 (2011).

143. Juhlin, K., Star, K. & Norén, G. N. A method for data-driven exploration to pinpoint key features in medical data and facilitate expert review. *Pharmacoepidemiology and Drug Safety* **26,** 1256–1265. doi:10.1002/pds.4285 (2017).

144. Cragg, J. G. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica* **39,** 829–844 (1971).

145. Cameron, A. C. & Trivedi, P. K. *Regression Analysis of Count Data* 2nd ed. ISBN: 9781139013567. doi:10.1017/CBO9781139013567 (Cambridge University Press, 2013).

146. Zucchini, W., MacDonald, I. L. & Langrock, R. *Hidden Markov Models for Time Series: An Introduction using R* 2nd ed. ISBN: 978-1482253832 (CRC Press, 2016).

147. Petersen, A. H., Osler, M. & Ekstrøm, C. T. Data-Driven Model Building for Life-Course Epidemiology. *American Journal of Epidemiology.* doi:10.1093/aje/kwab087 (Mar. 2021).

148. Hernán, M. A., Hsu, J. & Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE* **32,** 42–49. doi:10.1080/09332480.2019.1579578 (2019).

149. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* (Chapman & Hall/CRC, Boca Raton, 2020).

150. 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. doi:10.1093/aje/kwv254 (2016).

151. Kiureghian, A. D. & Ditlevsen, O. Aleatory or epistemic? Does it matter? *Structural Safety* **31,** 105–112. doi:https://doi.org/10.1016/j.strusafe.2008.06.020 (2009).

152. Rathnam, C., Lee, S. & Jiang, X. An algorithm for direct causal learning of influences on patient outcomes. *Artificial Intelligence in Medicine* **75,** 1–15. doi:10.1016/j.artmed.2016.10.003 (Jan. 2017).

153. Spirtes, P. & Glymour, C. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review* **9,** 62–72. doi:10.1177/089443939100900106 (Apr. 1991).

154. Kohane, I. S. *et al.* What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *Journal of Medical Internet Research* **23,** e22219. doi:10.2196/22219 (Mar. 2021).

155. Sundhedsdatastyrelsen. *Fællesindhold for basisregistrering af sygehuspatienter 2019-1: Vejledningsdel* tech. rep. (Sundhedsdatastyrelsen, 2018).

156. Stewart, W. F., Shah, N. R., Selna, M. J., Paulus, R. A. & Walker, J. M. Bridging the inferential gap: the electronic health record and clinical evidence. *Health Affairs (Millwood)* **26,** w181–91. doi:10.1377/hlthaff.26.2.w181 (Apr. 2007).

157. Drummond, M. F., Sculpher, M. J., Claxton, K. L., Stoddart, G. L. & Torrance, G. W. *Methods for the Economic Evauation of Health Care Programmes* 4th ed. ISBN: 978-0-19-966588 (Oxford University Press, 2015).

158. Jensen, A. B. *et al.* Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications* **5.** http://dx.doi.org/10.1038/ncomms5022 (June 2014).

159. Editorial. Revive clinical research. *Prescrire International* **28,** 171 (2019).

160. Christensen, S., Jensen, L. D., Kaae, S., Vinding, K. L. & Petersen, J. Implementation of the shared medication record is difficult [in Danish]. *Ugeskr Laeger* **176,** 1389–1391 (July 2014).

161. Thysen, S. M. & Ullum, H. *Debat: Forskere bør dele oprydningsnøgler for sundhedsdata [in Danish]* [Online; accessed 1. Nov. 2021]. https://www.altinget.dk/forskning/artikel/debat-forskere-boer-dele-oprydningsnoegler-for-sundhedsdata.

162. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3.** https://doi.org/10.1038/sdata.2016.18 (Mar. 2016).

163. *Dokumentation - Sygesikringsregisteret* [Online; accessed 7. Oct. 2021]. Oct. 2021. https://www.esundhed.dk/Dokumentation?rid=10.

164. *Dokumentation - Lægemiddelstatistikregistret* [Online; accessed 20 October 2021]. 2021. https://www.esundhed.dk/Dokumentation?rid=14.

165. *The Book of OHDSI [e-book]* 1st ed. https://ohdsi.github.io/TheBookOfOhdsi (Observational Health Data Sciences and Informatics, 2021).

166. Wilson, G. V. Where's the Real Bottleneck in Scientific Computing? *American Scientist* **94,** 5–6. https://www.proquest.com/scholarly-journals/wheres-real-bottleneck-scientific-computing/docview/215261934/se-2?accountid=13607 (Feb. 2006).

167. Morin, A *et al.* Research priorities. Shining light into black boxes. *Science* **336,** 159–160. doi:10.1126/science.1218263 (2012).

168. Dhombres, F., Winnenburg, R., Case, J. T. & Bodenreider, O. in *MEDINFO 2015: eHealth-enabled Health* 795–799 (IOS Press, 2015). doi:10.3233/978-1-61499-564-7-795.

169. Kaas-Hansen, B. S. *et al.* Hyponatraemia and mortality in psychiatric patients: protocol for Bayesian causal inference study. *medRxiv (unpublished).* doi:10.1101/2020.06.25.20138206 (2020).

170. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522. doi:10.1093/bioinformatics/bts480 (Aug. 2012).

171. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35,** 316–319. doi:10.1038/nbt.3820 (2017).

172. Norén, G. N., Sundberg, R., Bate, A. & Edwards, I. R. A statistical methodology for drug–drug interaction surveillance. *Statistics in Medicine* **27,** 3057–3070. doi:10.1002/sim.3247 (2008).

173. Institute of Medicine. *The Learning Healthcare System: Workshop Summary* (eds Leigh, A. O., Aisner, D. & McGinnis, J. M.) ISBN: 978-0-309-10300-8. doi:10.17226/11903 (The National Academies Press, Washington, DC, 2007).

174. Stang, P. E. *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine* **153,** 600–606. doi:10.7326/0003-4819-153-9-201011020-00010 (2010).

175. Kum, H.-C. & Ahalt, S. Privacy-by-Design: Understanding Data Access Models for Secondary Data. *AMIA Summits on Translational Science Proceedings* **2013,** 126–130 (2013).

176. Di Iorio, C. T. *et al.* Cross-border flow of health information: is 'privacy by design' enough? Privacy performance assessment in EUBIROD. *European Journal of Public Health* **23,** 247–253. doi:10.1093/eurpub/cks043 (Apr. 2013).

177. Hertzman, C. P., Meagher, N. & McGrail, K. M. Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. *Journal of the American Medical Informatics Association* **20,** 25–28. doi:10.1136/amiajnl-2012-001011 (Jan. 2013).

178. Johnson, A. E. W. *et al.* Machine Learning and Decision Support in Critical Care. eng. *Proc IEEE Inst Electr Electron Eng* **104,** 444–466. doi:10.1109/JPROC.2015.2501978 (2016).

179. Chrétien, B. *et al.* Disproportionality analysis in VigiBase as a drug repositioning method for the discovery of potentially useful drugs in Alzheimer's disease. *British Journal of Clinical Pharmacology* **87,** 2830–2837. doi:10.1111/bcp.14690 (July 2021).

180. Conway, D. *The Data Science Venn Diagram – Drew Conway* Online. [Online; accessed 21. Oct. 2021]. 2010. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

181. Alarcón-Soto, Y. *et al.* Data Science in Biomedicine. *arXiv (unpublished)* (2019).

182. Trifiro, G., Sultana, J. & Bate, A. From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources. *Drug Safety* **41,** 143–149. doi:10.1007/s40264-017-0592-4 (Feb. 2018).

183. Mack, C, Su, Z & Westreich, D. *Managing Missing Data in Patient Registries – Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition* tech. rep. (Agency for Healthcare Research and Quality, Rockville (MD), 2018). https://www.ncbi.nlm.nih.gov/books/NBK493611.

184. Committee for Human Medicinal Products. *Guideline on registry-based studies [EMA/426390/2021]* https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en-0.pdf (European Medicines Agency, 2021).

185. Trifirò, G. *et al.* Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? eng. *Pharmacoepidemiology and Drug Safety* **18,** 1176–1184. doi:10.1002/pds.1836 (2009).

186. *Key Database Statistics | Sentinel Initiative* [Online; accessed 28. Oct. 2021]. 2021. https://www.sentinelinitiative.org/about/key-database-statistics.

187. Klünemann, M. *et al.* Bioaccumulation of therapeutic drugs by human gut bacteria. *Nature* **597,** 533–538. doi:10.1038/s41586-021-03891-8 (2021).

188. Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R. & Goodman, A. L. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* **570,** 462–467. doi:10.1038/s41586-019-1291-3 (2019).

189. Mazurowski, M. A., Buda, M., Saha, A. & Bashir, M. R. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging* **49,** 939–954. doi:10.1002/jmri.26534 (2019).

190. Prahs, P. *et al.* OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications. *Graefe's Archive for Clinical and Experimental Ophthalmology* **256,** 91–98. doi:10.1007/s00417-017-3839-y (2018).

191. BenTaieb, A. & Hamarneh, G. Deep Learning Models for Digital Pathology. *arXiv (unpublished).* https://arxiv.org/abs/1910.12329 (2019).

192. Strickland, E. How IBM Watson Overpromised and Underdelivered on AI Health Care. *IEEE Spectrum.* https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care (2021).

193. Greene, T. J., DeSantis, S. M., Brown, D. W., Wilkinson, A. V. & Swartz, M. D. A machine learning compatible method for ordinal propensity score stratification and matching. *Statistics in Medicine* **40,** 1383–1399. doi:10.1002/sim.8846 (Mar. 2021).

194. Cannas, M. & Arpino, B. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal* **61,** 1049–1072. doi:10.1002/bimj.201800132 (July 2019).

195. Kalisch, M. & Bühlmann, P. Causal Structure Learning and Inference: A Selective Review. *Quality Technology & Quantitative Management* **11,** 3–21. doi:10.1080/16843703.2014.11673322 (2014).

196. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology* **45,** 1866–1886. doi:10.1093/ije/dyw314 (Jan. 2017).

Acknowledgements

First and foremost, I would like to thank my supervisors—Stig, Søren and Gesche—for their guidance and patience; for helping me grow as a scientist and mold my academic profile; and for letting me experience the equally exciting, frustrating, uplifting, and challenging world of big data and clinical research.

Thank you, Stig, for your valuable, sincere feedback and for giving me a strong base in Roskilde; Søren, for inviting me into you lab and helping me shape my idea of how I would manage a lab if I get the chance; and Gesche, you for our work on i.a. network meta-analyses and for always shrewd inputs and insights.

I am forever grateful for my office mates at the NNF Centre for Protein Research. Anna and HC patiently put up with my questions, doubts, and rants. Later, Cristina and Davide joined that club, and they have all become genuine friends who taught me much about medicine, bioinformatics, analytic programming, life, the universe, and everything.

The results of this thesis rest on hard work carried out by many people over time in Brunaklab. I have played an active role in some parts of data wrangling (some might call it wrestling), but in many respects I stand on the shoulders of giants who managed to mold the raw data and cast them into the best possible format. They all deserve praise.

I would like to thank Francesco Russo, David Westergaard, Peter C. Holm, and Peter Bruun for many interesting discussions on i.a. models, Bayes, science, databases, multiprocessing, and Snakemake.

Uppsala Monitoring Centre kindly hosted my change of environment and accommodated me and my family for a beautiful (if dark and cold) stay in Uppsala late-2019. Thank you—especially Niklas, Lucie, Oskar and the Data Science team—for letting me experience research outside the university walls. Unfortunately my planned repeated visits throughout 2020 never happened due to the COVID-19 pandemic, but I hope to get the chance to work with you guys again in the future.

Last but in no way least, I would like to thank my family: I am so grateful for your support. In particulare Simona, mia cara moglie e mamma dei nostri figli, Sofia e Sebastian. Sarò per sempre gratissimo per il tuo aiuto, il tuo sostegno, la tua pazienza e il tuo amore. Sono davvero fortunato! Senza di te, davvero, non ce l'avrei mai fatta—infatti non l'avrei neanche cominciato.