European Journal of Epidemiology
https://doi.org/10.1007/s10654-018-0386-8

PHARMACO-EPIDEMIOLOGY

Hypothesis-free screening of large administrative databases for unsuspected drug-outcome associations

Jesper Hallas[1,2] · Shirley V. Wang[1] · Joshua J. Gagne[1] · Sebastian Schneeweiss[1] · Nicole Pratt[3] · Anton Pottegård[2]

Received: 12 January 2018 / Accepted: 21 March 2018

Abstract
Active surveillance for unknown or unsuspected adverse drug effects may be carried out by applying epidemiological techniques to large administrative databases. Self-controlled designs, like the symmetry design, have the advantage over conventional design of adjusting for confounders that are stable over time. The aim of this paper was to describe the output of a comprehensive open-ended symmetry analysis of a large dataset. All drug dispensings and all secondary care contacts in Denmark during the period 1995–2012 for persons born before 1950 were analyzed by a symmetry design. We analyzed all drug–drug sequences and all drug–disease sequences occurring during the study period. The identified associations were ranked according to the number of outcomes that potentially could be attributed to the exposure. In the main analysis, 29,891,212 incident drug therapies, and 21,300,000 incident diagnoses were included. Out of 186,758 associations tested in the main analysis, 43,575 (23.3%) showed meaningful effect size. For the top 200 drug–drug associations, 47% represented unknown associations, 24% represented known adverse drug reactions, 30% were explained by mutual indication or reverse causation. For the top 200 drug–disease associations the proportions were 31, 15, and 55%, respectively. Screening by symmetry analysis can be a useful starting point for systematic pharmacovigilance activities if coupled with a systematic post-hoc review of signals.

Keywords Pharmacovigillance · Pharmcoepidemiology · Self-controlled design · Databases · Screening

Background

About 1–3% of all newly marketed drugs are withdrawn because of adverse effects that are not known at the time of authorization [1, 2], thus necessitating a systematic surveillance of marketed drugs. For decades, the primary tool in generating signals about adverse drug effects after marketing has been spontaneous reporting [3]. This approach has several well-known limitations. First, there is massive and, most importantly, highly variable underreporting [4, 5], which is often the underlying cause of signals in spontaneous reporting schemes [6]. Second, individual case reports require an individual patient or clinician to connect the drug and the adverse event as potentially being causally related. It is therefore likely that many inconspicuous adverse drug reactions (e.g., those involving common events or with insidious onsets) are never captured by spontaneous reporting. Third, the spontaneous reporting scheme is highly sensitive to media attention, in that controversy surrounding an adverse drug effect might in itself generate a surge in reports. These limitations—and possibly others—might be addressed by a systematic, open-ended epidemiological analysis of large administrative databases [7], in which associations between drug use and outcomes, as they occur in clinical practice, are assessed without relying on reporting.

✉ Jesper Hallas
jhallas@health.sdu.dk

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[2] Clinical Pharmacology and Pharmacy, Department of Public Health, University of Southern Denmark, JB Winsløwsvej 19, 2, 5000 Odense C, Denmark

[3] Quality Use of Medicines and Pharmacy Research Centre, Sansom Institute, University of South Australia, Adelaide, Australia

Published online: 31 March 2018

One challenge to analysing administrative databases for screening purposes—at least with conventional cohort and case-control designs—is that it takes considerable thought, time and effort to design these studies and choose appropriate comparators in a way that minimizes confounding and generation of false positive results. As it would be infeasible to meticulously select relevant comparators and confounders for each potential association based on subject matter knowledge, study designs that do not require identification of an external comparator group and a formal adjustment for confounders are desirable in a hypothesis-free screening system. Self-controlled designs rely on within-person comparisons, which provides the advantage of being robust towards confounders that are stable over time, and some are simple enough to be amenable to large-scale automated screening [8].

The symmetry analysis may be a particularly useful study design for early stage hypothesis-free screening for adverse drug effects. Its processing is very simple [9], it has reasonable sensitivity and specificity [10], and it has the key properties of a self-controlled design [11]. To our knowledge, all published papers on this method have focussed on a limited number of outcomes or exposures. The aim of this paper was to describe its yield in an open-ended hypothesis generating screening of a large dataset. This work appraises its potential value as a first step in active surveillance.

Methods

We analysed all prescriptions between 1995 and 2012 and all secondary care contacts between 1994 and 2012 for residents of Denmark born 1950 or earlier, using a symmetry design.

Setting

Three nation-wide Danish data sources were used: The Danish National Prescription Registry [12], the Danish Patient Registry [13], and the Civil Registration System [14].

Virtually all secondary care in Denmark is provided by the national health authorities, allowing true population-based register linkage studies covering all inhabitants of Denmark. Data were linked by the personal identification number, a unique identifier assigned to all Danish residents since 1968. All linkages were performed within Statistics Denmark, a governmental institution that collects and processes information for a variety of statistical and scientific purposes [15].

The project was approved by the scientific review board of the Danish National Health Board. According to Danish law, approval from an ethics committee is not required for pure register studies [15].

Data

The data employed in this analysis included all Danish residents born 1950 or earlier who had at least one prescription recorded during the period 1995–2012 (N = 1,848,825). Of these, 1,796,446 (97%) also had a secondary care diagnosis recorded. In all, 479,420,576 prescriptions and 80,865,480 secondary care diagnoses were included in the analysis.

Design

The symmetry design was first described by Hallas in 1996 in a study on depression provoked by cardiovascular medication [9]. If, for example, we sought to assess whether thiazides might cause depression, we would identify all persons who, at different dates within a defined interval (e.g. 12 months), initiate therapy with both thiazides and antidepressants for the first time in their lives. In this particular population, and assuming no association between the two, we would expect a symmetrical distribution of persons starting either drug first; i.e., as many who started thiazides before antidepressants as persons who followed the opposite order. However, if thiazides cause depression, there would be a relative excess of persons starting thiazides first and then antidepressants. It can be shown that the sequence ratio, i.e. the ratio of counts of persons who start thiazides first versus those who start antidepressants first, is an estimate of the incidence rate ratio of antidepressant prescribing in follow-up exposed versus non-exposed to thiazides, possibly with a small conservative bias [16].

We used the variant of the symmetry analysis, where the exposure drug is anchored in time, and the rate of the outcome is estimated in a symmetrical time-window before and after the first prescription of the exposure drug [17]. This specific symmetry design variant is not biased by temporal trends in use of exposure drugs, although trends in the outcome drug or event may confer a bias if not appropriately adjusted [17]. The width of the interval was set to 12 months on either side of the index date, and sensitivity analyses were performed by setting the width of the interval to 6 or 18 months. We used dispensing of drugs as the exposure in all analyses. However, we used both secondary care diagnoses and dispensing of drugs to represent outcomes. In keeping with the thiazide-depression example above, we would thus analyze the distribution of new antidepressant treatments before and after the first thiazide prescriptions, but also the distribution of new secondary care contacts with a diagnosis of depression.

Analysis

The entire history of each individual was used to establish first occurrence of a prescription or a diagnosis—i.e., a person who had a single prescription for a beta-blocker in 2002 and again in 2009, was only counted as an incident beta-blocker user in 2002. To allow for the proper recognition of incident therapies and diagnoses, we refrained from including first diagnoses or treatments occurring before January 1996, thus ensuring a run-in period of at least 1 year for prescription data and 2 years for secondary care contacts.

Both the ATC, used to categorize drugs, and the ICD10, used to categorize diagnoses, are hierarchical classification systems. By considering an increasing number of digits of the code, increasing precision is achieved. For example, M01A indicates the ATC code for all NSAIDs, M01AE for all propionic acid derivatives and M01AE01 for ibuprofen. In our main analysis, we employed four digits of the ATC code (corresponding to its 3rd level) and three digits of the ICD10 code. In order to be able to further elucidate potential signals, we also analysed all codes according to the 5th and 7th digit of the ATC code (4th and 5th level, respectively) and to the full four-digit ICD10 code.

In addition, we employed the following rules to limit the number of signals:

- We only reported association with sequence ratios above unity, indicating potential harmful drug effects.
- We removed drug–drug pairs in which the drugs shared the first digit of the ATC code. Such pairs were considered likely to represent drugs prescribed for the same indication where the sequence merely reflects that first-line drugs are typically prescribed before second-line drugs.
- We removed drug–disease pairs where the first digit of the ATC codes and of the ICD10 codes suggested that they belonged to the same organ system. For example, pairs consisting of a cardiovascular drug and a cardiovascular diagnosis were removed, as these were considered likely to represent confounding by indication. A list of these exclusions is shown in Appendix 1.
- We only considered exposure drugs or drug classes with more than 10,000 incident users.

The impact of each of these measures on the number of potential outcome codes and the number of potential signals was analysed and described. The main analysis was based on 4-digit ATC codes and 3-digit ICD10 codes, a 12-month time window and with implementation of all the above limiting measures.

Due to the hypothesis generating nature of the study, we did not adjust for multiple testing [18], and we did not adjust for trends in the individual outcomes [9], due to the overwhelming computational time requirement. The results are thus presented as crude sequence ratios, unadjusted for time trends. Confidence intervals for sequence ratios were calculated by use of the methods described by Morris and Gardner [19], and p values for the sequence ratios were calculated by simple Chi square tests.

We ranked the output according to the absolute numerical difference in sequence orders. If, for example, drug A preceded B 4000 times and B preceded A 1000 times, then the numerical difference is 3000. The 30 signals with the highest numerical differences are shown in Table 3 (drug–drug pairs) and Table 4 (drug–disease pairs). The underlying rationale for choosing this ranking was that—all other things being equal—this would represent the signals with the potentially highest public health impact.

The 200 highest-ranking findings in each category (drug–drug and drug–disease associations) were classified into broad categories:

- *ADR*, Known adverse drug reactions.
- *RC*, Reverse causation, i.e., that the drug is used to treat an early manifestation of a condition that later becomes clinically apparent or is later diagnosed in secondary care. For example, a drug for overactive bladder may be prescribed for symptomatic relief for a patient complaining of urge and frequent voiding. Upon further work-up, the patient may be found to have bladder cancer and is given a bladder cancer diagnosis. This creates a reverse-causation link between drugs for overactive bladder and bladder cancer.
- *MIC*, Mutual indications or causes—e.g. there could be asymmetry between paracetamol and opioids, since they are both prescribed for pain, but paracetamol usually before opioids.
- *TDC*, Time dependent confounding—e.g., paracetamol is prescribed together with an opioid in some patients. If the opioid causes constipation, it will lead to asymmetry between paracetamol and laxatives. Opioid use is a confounder of this association, which, since it is temporally linked to paracetamol use, is not inherently controlled by the symmetry design.
- *Unknown* Not readily explainable. Possibly an unknown adverse drug reaction.

This categorisation was performed independently by two of the authors (JH and AP), and disagreements were resolved by consensus.

Results

Out of the 1,851,352 drug users analysed, 858,904 (46%) were men and 992,448 (54%) were women. The median age at their first recorded prescription was 58 years with an interquartile range of 51–68.

Outcome metrics

We analysed 29,891,212 different treatments, i.e. instances of unique drug use (specified to the third level of the ATC system) in unique individuals and 21,300,000 disease episodes, i.e. unique disease events (specified to the third digit of the ICD10-code) in unique individuals. When the fully specified ATC and ICD10 codes were used, 44,008,672 treatments and 24,117,268 disease episodes were analysed (Table 1). The number of individual pairs (drug–drug or drug–disease sequences in individual persons) included in the main analysis was 3.8×10^{10}, and 23×10^{11} when the full codes were used with no restrictions of pairs.

The impact of various measures to restrict the analyses is shown in Table 2. The largest impact was achieved by using truncated ATC- and ICD codes, from 3,099,493 different associations to 244,891. The effects of other restrictions were fairly modest; the main analysis encompassed 186,758 different associations (Table 2).

Top ranking association

The highest ranking drug–drug association was NSAID → opioid (sequence ratio (SR) 2.14, 95% confidence interval (CI) 2.11–2.16) (Table 3). We interpreted it as a mutual indication, pain, which would typically be treated with NSAIDs before opioids were attempted. The second highest ranking drug–drug association was opioids → laxatives (SR 2.34, CI 2.31–2.38), which we interpreted as reflecting a well-known ADR from opioids, constipation. The highest ranking drug–disease association was anti-ulcer drugs → dyspepsia (SR 2.46, CI 2.39–2.53) (Table 4). We interpreted it as an example of reverse causation, reflecting that anti-ulcer drugs are prescribed with little delay in primary care, while secondary care diagnostic work-up usually entails a certain delay. We interpreted the twelve highest ranking drug–disease associations as variants of reverse causation. The highest ranking association which was interpreted as reflecting an ADR was corticosteroids → osteoporosis (SR 2.20, CI 2.10–2.30).

The top 30 drug–drug analyses revealed five associations that may require further examination: antithrombotics → drugs against constipation; antithrombotics → minor analgesics; antithrombotics → antidepressants; beta-blockers → minor analgesics; and NSAIDs → cough suppressants. Similarly, the drug–disease analyses produced two unexplained associations, opioids → dehydration and NSAIDs → pneumonia. All top 30 drug–drug or drug–disease associations had p values below 10^{-83}.

The distribution of interpretations for the 200 highest-ranking drug–drug and drug–disease associations are shown in Table 5. The dominant interpretations of drug–drug association were mutual indications or causes (16%) or "unknown" (53%), whereas the dominant interpretations of drug–disease associations were reverse causation (50%) and unknown (34%).

Discussion

In our screening, we were able to reproduce a large number of known ADRs and a large number of associations that reflect everyday sound clinical behaviour, such as prescribing first-line before second-line drugs. In addition, a number of associations are readily interpretable as time-dependent confounding, a reverse causation or a result of a mutual indications or causes. Some of the remaining associations, denoted as unknown causes, may well fall into the same categories after further evaluation, or, in a small fraction of those, may hypothesize about unsuspected ADRs. It is vital to emphasize that this is a first, crude screening and that an extensive follow-up of these signals is warranted before any inferences or conclusions should be

Table 1 Data material included in the hypothesis-free symmetry screening of administrative Danish data sources

Data source	Individual patients	Individual records	Different patient-level codes[a], full code	Different patient-level codes[a], truncated code
Prescriptions	1,848,825	479,420,576	44,008,672	29,891,212
Secondary care diagnoses	1,796,446	80,865,480	24,117,268	21,300,000

ATC Anatomic Therapeutic Chemical classification system, *ICD10* International Classification of Diseases, version 10

[a]A patient-level code refers to a specific diagnosis or prescription occurring at least once in a patient. The codes used in the main analysis are truncated to the four digits for the drug code (ATC) and three digits for the disease codes (ICD10)

Table 2 Outcome metrics for comprehensive symmetry analysis of all prescriptions and secondary contacts in Denmark since 1995 for persons born 1950 or earlier

Number of digits in ATC codes	Number of digits in ICD10 code	Time window allowed between index prescription and out-come, months	Other criteria	Number of individual-level drug–drug and drug–disease pairs analyzed	Number of different potential associations analyzed
7	4	12	None	232,303,361,934	3,099,493
5	4	12	None	194,871,416,068	1,509,189
4	4	12	None	153,848,455,003	777,213
7	3	12	None	84,805,221,817	1,485,156
5	3	12	None	57,765,050,448	578,946
4	3	12	None	40,141,896,088	244,891
4	3	6	None	38,801,361,260	227,034
4	3	18	None	40,887,571,331	255,174
4	3	12	No drug–drug pairs where first digit in ATC-code is identical	39,720,482,581	241,854
4	3	12	No drug–disease pairs where drug and disease belong to the same organ system	38,725,301,455	235,720
4	3	12	At least 10,000 treatments for the given drug	39,982,840,125	195,873
4	3	12	All of the last three (main analysis)	38,154,194,500	186,758
5	4	12	None	194,871,416,068	1,509,189
4	4	12	None	153,848,455,003	777,213
7	3	12	None	84,805,221,817	1,485,156
5	3	12	None	57,765,050,448	578,946
4	3	12	None	40,141,896,088	244,891

The table shows the size of the data material, the number of different pairs to analyze and the proportion of statistically significant sequence ratios, given various criteria for inclusion of drug–drug or drug–disease association

ATC Anatomic Therapeutic Chemical classification system, *ICD10* International Classification of Diseases, version 10

made about specific associations [20]. This post-hoc assessment could include such elements as a thorough mapping of the timing of events [11, 21], analyses by active comparator designs using the same data, analyses with adjustment for selected confounders, analyses of effect modifications [9, 22], assessment of mechanistic plausibility [23], analyses for dose-response effect, formal assessment by a panel of clinical experts, and analyses in different data set using the same approach [24]. The optimal combination of data analytic approaches of a comprehensive screening system is yet to be established, but our results suggest that open-ended screening using symmetry analyses may be a useful early step in the process.

A key limitation of open-ended screening in electronic healthcare data is that, in contrast to spontaneous reporting, there is no built-in "clinical filter" to weed out non-informative associations For example, clinicians would be unlikely to report opioid-induced constipation to regulatory agencies as this would usually be considered trivial. Likewise, clinicians would not report that NSAIDs were prescribed before opioids in a patient, as this represents sound use of the WHO pain ladder [25] rather than a suspected ADR. These sequences are extremely common in clinical practice and thus appear among the highest-ranking signals in our analysis. Not surprisingly, reverse causation explained a substantial proportion of drug–disease associations, while this mechanism was rarely seen with drug–drug associations. We interpret this as a consequence of the fact that the diagnoses we had were exclusively from secondary care and of the inevitable delay in referral and diagnostic work-up in this setting. For example, a patient may have had lumbar pain for some time before being referred to a hospital specialist and have his diagnosis recorded. Analgesics are prescribed without this delay, thus producing a reverse signal of analgesics → lumbar pain. This step of identifying associations that have clear non-causal explanations was critical, much

Table 3 Top 30 signals in main analysis, ranked in descending order by difference in number of patients with exposure drug first versus exposure drug second. Drug–drug pairs only

Rank	Exposure drug (ATC)	Outcome drug (ATC)	Exposure drug first/last	Time difference, days, median (IQR)[a]	Sequence ratio (95% confidence interval)	Log10(*p*)	Interpretation
1	NSAIDs (M01A)	Opioids (N02A)	73,793/34,542	59 (12–186)	2.14 (2.11–2.16)	< − 300	MIC
2	Opioids (N02A)	Drugs for constipation (A06A)	58,893/25,127	66 (21–173)	2.34 (2.31–2.38)	< − 300	ADR
3	Antithrombotic agents (B01A)	Cholesterol lowering drugs (C10A)	58,859/29,759	91 (31–190)	1.98 (1.95–2.01)	< − 300	MIC
4	High-ceiling diuretics (C03C)	Potassium (A12B)	37,271/11,268	43 (13–135)	3.31 (3.24–3.38)	< − 300	ADR
5	Thiazides (C03A)	Potassium (A12B)	35,047/10,128	105 (35–220)	3.46 (3.39–3.54)	< − 300	ADR
6	Opioids (N02A)	Propulsives (A03F)	42,322/19,817	63 (21–168)	2.14 (2.10–2.17)	< − 300	ADR
7	Other analgesics (= Paracetamol) (N02B)	Drugs for constipation (A06A)	48,385/28,207	98 (33–212)	1.72 (1.69–1.74)	< − 300	TDC
8	NSAIDs (M01A)	Other analgesics (= Paracetamol) (N02B)	45,941/27,478	99 (31–217)	1.67 (1.65–1.70)	< − 300	MIC
9	NSAIDs (M01A)	Corticosteroids for systemic use, plain (H02A)	36,346/19,479	87 (28–205)	1.87 (1.83–1.90)	< − 300	MIC
10	Opioids (N02A)	Drugs affecting bone structure and mineralization (M05B)	20,776/5484	95 (42–190)	3.79 (3.68–3.90)	< − 300	MIC
11	Potassium (A12B)	Potassium-sparing agents (C03D)	21,464/6472	97 (36–204)	3.32 (3.23–3.41)	< − 300	MIC
12	Antithrombotic agents (B01A)	Drugs for constipation (A06A)	26,205/12,348	134 (51–241)	2.12 (2.08–2.17)	< − 300	Unknown
13	NSAIDs (M01A)	Anti-ulcer drugs (A02B)	31,295/18,351	136 (50–246)	1.71 (1.67–1.74)	< − 300	ADR
14	Antithrombotic agents (B01A)[b]	Other analgesics (= Paracetamol) (N02B)	40,143/27,300	138 (56–244)	1.47 (1.45–1.49)	< − 300	Unknown
15	NSAIDs (M01A)	Drugs for constipation (A06A)	20,841/9888	126 (49–240)	2.11 (2.06–2.16)	< − 300	TDC
16	Opioids (N02A)	Antacids (A02A)	21,060/10,818	92 (31–203)	1.95 (1.90–1.99)	< − 300	TDC
17	Antithrombotic agents (B01A)[b]	Antidepressants (N06A)	27,913/17,906	135 (56–239)	1.56 (1.53–1.59)	< − 300	Unknown
18	Antithrombotic agents (B01A)[b]	Anti-ulcer drugs (A02B)	32,161/22,780	140 (57–245)	1.41 (1.39–1.44)	< − 300	ADR
19	Other analgesics (= Paracetamol) (N02B)	Antacids (A02A)	20,857/11,708	116 (45–226)	1.78 (1.74–1.82)	< − 300	TDC
20	Inhaled beta-agonists (R03A)	Corticosteroids for systemic use, plain (H02A)	22,579/13,437	108 (35–226)	1.68 (1.64–1.72)	< − 300	MIC
21	Other analgesics (= Paracetamol) (N02B)	Propulsives (A03F)	27,228/18,550	90 (29–205)	1.47 (1.44–1.50)	< − 300	TDC
22	Cough suppressants (R05D)	Drugs for constipation (A06A)	17,189/8826	105 (36–219)	1.95 (1.90–2.00)	< − 300	ADR
23	Beta blockers (C07A)	Other analgesics (= Paracetamol) (N02B)	23,324/15,477	141 (58–243)	1.51 (1.48–1.54)	< − 300	Unknown
24	Beta-lactam antibacterials, penicillins (J01C)	Agents against amoebiasis and other protozoal diseases (P01A)	11,845/4241	88 (17–222)	2.79 (2.70–2.89)	< − 300	MIC
25	NSAIDs (M01A)	Propulsives (A03F)	16,674/9156	124 (47–235)	1.82 (1.78–1.87)	< − 300	TDC
26	Macrolides, lincosamides and streptogramins (J01F)	Inhaled anticholinergics and steroids (R03B)	160,05/8585	90 (29–212)	1.86 (1.82–1.91)	< − 300	MIC
27	NSAIDs (M01A)	Cough suppressants (R05D)[c]	22,650/15,292	105 (25–227)	1.48 (1.45–1.51)	< − 300	MIC

Table 3 (continued)

Rank	Exposure drug (ATC)	Outcome drug (ATC)	Exposure drug first/last	Time difference, days, median (IQR)[a]	Sequence ratio (95% confidence interval)	Log10(*p*)	Interpretation
28	Beta blockers (C07A)	Antidepressants (N06A)	18,277/11,088	138 (57–242)	1.65 (1.61–1.69)	< − 300	Unknown
29	Corticosteroids for systemic use, plain (H02A)	Drugs affecting bone structure and mineralization (M05B)	10,064/2959	159 (83–252)	3.40 (3.27–3.54)	< − 300	ADR
30	Antithrombotic agents (B01A)	Calcium channel blockers (C08C)	32,260/25,173	115 (46–221)	1.28 (1.26–1.30)	− 191.8	MIC

ATC Anatomic Therapeutic Chemical classification system, *IQR* interquartile range, *Log10(p)* Base 10 logarithm of *p* value, *MIC* mutual indication or cause, *TDC* time dependent confounding, *ADR* known adverse drug reaction

[a]Time interval in days between exposure and outcome drugs for sequences following this sequence

[b]Does not include aspirin in analgesic doses

[c]Includes codeine tablets

Table 4 Top 30 signals in main analysis, ranked according to falling difference in numbers having exposure drug first versus last. Drug–disease pairs only

Rank	Exposure drug (ATC)	Outcome (ICD10)	Exposure drug first/last	Time difference, days, median (IQR)[a]	Sequence ratio (95% confidence interval)	Log10(*p*)	Interpretation
1	Anti-ulcer drugs (A02B)	Dyspepsia (K30)	17,487/7115	74 (36–154)	2.46 (2.39–2.53)	< − 300	RC
2	Opioids (N02A)	Other intervertebral disc disorders (M51)	14,985/5552	56 (20–135)	2.70 (2.62–2.78)	< − 300	RC
3	Opioids (N02A)	Dorsalgia (M54)	16,670/9858	55 (15–156)	1.69 (1.65–1.73)	< − 300	RC
4	Opioids (N02A)	Osteoporosis without pathological fracture (M81)	11,804/5563	105 (44–210)	2.12 (2.06–2.19)	< − 300	RC
5	Corticosteroids for systemic use, plain (H02A)	Shoulder lesions (M75)	8084/2297	134 (76–216)	3.52 (3.36–3.69)	< − 300	RC
6	Macrolides, lincosamides and streptogramins (J01F)	Pneumonia, organism unspecified (J18)	14,603/9984	75 (10–211)	1.46 (1.43–1.50)	− 190.3	RC
7	Macrolides, lincosamides and streptogramins (J01F)	Other chronic obstructive pulmonary disease (J44)	10,277/6067	110 (32–229)	1.69 (1.64–1.75)	− 237.4	RC
8	Topical treatment of hemorrhoids and anal fissures (C05A)	Other diseases of anus and rectum (K62)	6787/2695	67 (27–168)	2.52 (2.41–2.63)	< − 300	RC
9	Anti-ulcer drugs (A02B)	Cholelithiasis (K80)	9390/5554	91 (35–191)	1.69 (1.64–1.75)	− 215.7	RC
10	Beta-lactam antibacterials, penicillins (J01C)	Pneumonia, organism unspecified (J18)	15,473/11,721	82 (12–219)	1.32 (1.29–1.35)	− 114.2	RC

Table 4 (continued)

Rank	Exposure drug (ATC)	Outcome (ICD10)	Exposure drug first/last	Time difference, days, median (IQR)[a]	Sequence ratio (95% confidence interval)	Log10(*p*)	Interpretation
11	Opioids (N02A)	Other spondylopathies (M48)	7639/4203	98 (39–200)	1.82 (1.75–1.89)	− 218.4	RC
12	Opioids (N02A)	Other functional intestinal disorders (K59)	9301/5982	93 (28–212)	1.55 (1.51–1.61)	− 158.3	RC
13	Corticosteroids for systemic use, plain (H02A)	Osteoporosis without pathological fracture (M81)	6026/2742	146 (63–244)	2.20 (2.10–2.30)	− 269.0	ADR
14	Opioids (N02A)	Sequelae of injuries of lower limb (T93)	5994/2774	132 (53–231)	2.16 (2.07–2.26)	− 258.7	RC
15	NSAIDs (M01A)	Pneumonia, organism unspecified (J18)	9443/6225	147 (55–252)	1.52 (1.47–1.57)	− 145.3	Unknown
16	Macrolides, lincosamides and streptogramins (J01F)	Heart failure (I50)	8354/5152	119 (35–234)	1.62 (1.57–1.68)	− 166.7	Unknown
17	Macrolides, lincosamides and streptogramins (J01F)	Malignant neoplasm of bronchus and lung (C34)	4604/1440	60 (25–153)	3.20 (3.02–3.39)	< − 300	RC
18	Antithrombotic agents (B01A)	Other anaemias (D64)	8124/5072	147 (62–251)	1.60 (1.55–1.66)	− 155.1	ADR
19	NSAIDs (M01A)	Essential (primary) hypertension (I10)	13,825/10,774	156 (64–259)	1.28 (1.25–1.32)	− 83.9	ADR
20	Beta-lactam antibacterials, penicillins (J01C)	Other chronic obstructive pulmonary disease (J44)	7943/4904	114 (33–237)	1.62 (1.56–1.68)	− 157.9	RC
21	Vasodilators used in cardiac diseases (C01D)	Disorders of lipoprotein metabolism and other lipidaemias (E78)	13,154/10,136	82 (33–165)	1.30 (1.26–1.33)	− 86.6	MIC
22	Opioids (N02A)	Secondary malignant neoplasm of other and unspecified sites (C79)	7868/4934	72 (28–173)	1.59 (1.54–1.65)	− 147.8	RC
23	Opioids (N02A)	Complications of internal orthopaedic prosthetic devices, implants and grafts (T84)	6053/3129	110 (34–225)	1.93 (1.85–2.02)	− 204.1	RC
24	Opioids (N02A)	Volume depletion (E86)	11,039/8203	84 (26–197)	1.35 (1.31–1.38)	− 92.5	Unknown
25	NSAIDs (M01A)	Other anaemias (D64)	5205/2463	139 (52–246)	2.11 (2.02–2.22)	− 214.8	ADR
26	Opioids (N02A)	Osteoporosis with pathological fracture (M80)	6347/3623	72 (23–169)	1.75 (1.68–1.83)	− 163.4	RC
27	Beta-lactam antibacterials, penicillins (J01C)	Heart failure (I50)	9956/7318	119 (34–238)	1.36 (1.32–1.40)	− 89.2	Unknown
28	Opioids (N02A)	Other sepsis (A41)	6622/3991	101 (33–214)	1.66 (1.60–1.73)	− 143.4	Unknown
29	Anti-ulcer drugs (A02B)	Diaphragmatic hernia (K44)	7851/5278	72 (34–155)	1.49 (1.44–1.54)	− 111.2	RC
30	NSAIDs (M01A)	Cholelithiasis (K80)	6745/4184	82 (33–187)	1.61 (1.55–1.68)	− 132.1	RC

ATC Anatomic Therapeutic Chemical classification system, *ICD10* International Classification of Diseases, version 10, *Log10(p)* base 10 logarithm of *p* value, *MIC* mutual indication or cause, *ADR* known adverse drug reaction, *RC* reverse causation

[a]Time interval in days between exposure and outcome drugs for sequences following this sequence

Table 5 Distribution of signal causes in top 200 drug–drug pairs and drug–disease associations

Cause	Drug–drug associations	Drug–disease associations
Adverse drug reactions	31 (16%)	27 (14%)
Mutual indication or cause	52 (26%)	5 (3%)
Reverse causation	0 (0%)	100 (50%)
Time dependent confounding	11 (6%)	0 (0%)
Unknown	106 (53%)	68 (34%)

like the "clinical filter" in spontaneous adverse event reports.

A number of modifications of our screening approach need to be considered. First, we have focussed only on the first known use of specific prescription drugs in the patients. It is possible that allowing multiple distinct treatment episodes for the same drug in the same individual would enable analyses of more drug–drug or drug–disease pairs and thereby improve precision of our estimates.

Second, one may consider routinely adjusting for bias by trends in the outcome events. If, for example, the outcome drug shows an increasing trend over time, this may bias the symmetry ratio upward [9]. The "null-effect" approach can effectively adjust for this trend bias [16], but is quite demanding in terms of data requirements and processing. We could adjust for trends for the positive findings alone, but this would not alleviate the problem of signals that are overlooked because they are masked by a trend bias in the opposite direction.

Third, many of the produced signals represent already known adverse drug reactions. If these could be removed, for example by applying a digital library of known ADRs, it would substantially reduce the number of signals to be evaluated further. Unfortunately, for the drug–drug associations in our study, the outcomes are represented by drugs and not by clinical entities. We know of no data source of known ADRs where adverse drug reactions are systematically represented by other drugs potentially useful to treat them, but this is an opportunity for future enhancement.

Fourth, one might consider adjustment for multiple comparisons. If a simple criterion of statistical significance was applied using the conventional $p < 0.05$ threshold, our analysis generated 43,575 signals for further evaluation. It is obviously impossible to evaluate all of them in depth. However, simply lowering the significance level (e.g. to 0.001) is not a satisfactory solution. While it would reduce the number of false positive signals, it would also reduce the number of true positive signals, since they would now have to fulfil a more strict significance criterion. Thereby, lowering the significance threshold has little impact on the signal : noise ratio. In addition, given the rather extreme p values in the top 30 associations (Tables 3, 4), the issue is not how to handle chance findings, but rather how to separate trivial associations from those of interest.

Fifth, we prioritised signals on a simple criterion of numerical difference. In the absence of bias, this measure represents the number of outcomes attributable to the drug exposure [26], and thereby the largest public health impact, all other things being equal. This could be further refined by weighting the outcomes according to their severity, for example by giving more weight to a potentially drug-induced gastrointestinal bleeding than to drug-induced constipation, and then prioritise on the basis of weighted outcomes. This is, however, not a simple task, as these weights are subjective.

Sixth, one may consider applying this screening specifically for the first few years of marketing of a drug. As the experience with the drug grows, it may affect clinical behaviour and thereby the performance of the screening. For example, if a drug is wrongly suspected of causing depression, it will be less likely to be prescribed for persons with a history of depression. Thus, some pairs with the depression → drug sequence are avoided and a spurious signal of a drug-depression association would emerge [9]. The symmetry screening may therefore work most reliably with completely unsuspected associations and, probably, early in the market life of a drug. One cold also argue that early detection of unsuspected ADRs is by far the potentially most useful application of such screening.

Seventh, we chose to leave out associations arising from drugs within the same main therapeutic class and drug–disease pairs belonging to the same organ system, in order to limit the number of high-ranking non-causal signals. The disadvantage of such an approach is that some important ADRs are overlooked, e.g. extrapyramidal side effects of neuroleptics. With more experience, it is possible that an approach could be developed that allows for symmetry analyses within the same drug classes and organ systems.

Finally, signals should be validated in other data sources, whenever possible. Primary signals that emerge because of chance or a site-specific set of confounders are unlikely to be validated in other data sources. One example is the link between anti-epileptics and infections identified

in symmetry screening of data from Denmark [17], which could not be reproduced in a Taiwanese setting [24].

In conclusion, a comprehensive open-ended symmetry analysis produces mostly clinically interpretable results. Some findings are not readily interpretable. While the vast majority are likely explained by clinical behaviour, a few might represent unsuspected adverse drug effects. Although mining in large data sources has been practiced for several decades, nearly all pharmacovigillance findings are still based on spontaneous reporting schemes or vigillant clinical observers. The data mining described in our paper requires a high level of data infrastructure and would be feasible few places in the world. We do not suggest that data mining should replace conventional pharmacovigillance, Instead, open-ended data mining in large data sets could be a useful pharmacovigillance tool, when coupled with a systematic process to conduct post-hoc clinical assessment of signals and used as part of a holistic screening system that would also include spontaneous reporting. The most productive ways to conduct this are yet to be established.

Acknowledgements Funded by Clinical Pharmacology and Pharmacy, University of Southern Denmark, Denmark and Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, MA, USA.

Author contribution JH: Conceived the study, analyzed the data and wrote first draft. SVW, JJG, SS: Conceived the study, provided input to analysis and report. Nicole Pratt: Provided input to analysis and report. Anton Pottegård: Conceived the study, provided input to the report.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Appendix 1

Drug and disease main groups suggestive of confounding by indication. See text for explanation

Drug main group (ATC)	Drug group, plain text	Disease main group (ICD10)	Disease group, plain text
C	Cardiovascular system	I	Diseases of circulatory system
D	Dermatologicals	L	Skin diseases
G	Genito urinary system and sex hormones	N	Genitourinari and renal diseases
H	Systemic hormonal preparations, excl. sex hormones and insulins	E	Endocrine, nutritional and metabolic diseases
J	Antiinfectives for systemic use	B	Bacterial and certain viral infections
J	Antiinfectives for systemic use	A	Viral, fungal and parasitic infections
L	Antineoplastic and immunomodulating agents	C	Malignant neoplasms
L	Antineoplastic and immunomodulating agents	D	Benign neoplasms and hematological diseases
L	Antineoplastic and immunomodulating agents	M	Diseases of the musculoskeletal system and connective tissue
M	Musculo-skeletal system	M	Diseases of the musculoskeletal system and connective tissue
P	Antiparasitic products, insecticides and repellents	B	Viral, fungal and parasitic infections
P	Antiparasitic products, insecticides and repellents	A	Bacterial and certain viral infections
R	Respiratory system	J	Diseases of respiratory system
S	Sensory organs	H	Diseases of ear and eye

ATC Anatomical Therapeutic Chemical classification, *ICD10* International classification of diseases version 10

Appendix 2

Short description of the data sources included in the study.

The Danish National Prescription Registry contains data on all prescription drugs dispensed to Danish citizens from community pharmacies since 1995. Among other variables, the data include the dispensed substance, the date of dispensing, and quantity dispensed. Dosing information and indication for prescribing are not systematically recorded and were not used for this study. Drugs are categorized according to the Anatomic Therapeutic Chemical (ATC) index, a hierarchical classification system developed by the WHO, and the quantity dispensed for each prescription is given by the number of units and strength of the pharmaceutical product, as well as quantity expressed in the defined daily doses (DDD). The Danish National Prescription Registry does not include medications dispensed during hospitalization.

The Danish National Patient Register contains nationwide data on all non-psychiatric hospital admissions since 1977 and both psychiatric and non-psychiatric outpatient encounters since 1995. Discharge/contact diagnoses have been coded according to ICD-10 since 1994. Diagnoses established in primary care alone, i.e. without any involvement of hospital care, are not captured by the Danish National Patient Register.

The Danish Civil Registration System contains data on date of death and migrations to and from Denmark since 1968, which allowed us to keep track of all subjects during the study period.

References

1. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions. Arch Intern Med. 2005;165:1363–9.
2. Bakke OM, Manocchia M, de Abajo F, Kaitin KI, Lasagna L. Drug safety discontinuations in the United Kingdom, the United States, and Spain from 1974 through 1993: a regulatory perspective. Clin Pharmacol Ther. 1995;58:108–17.
3. Waller PC. Making the most of spontaneous adverse drug reaction reporting. Basic Clin Pharmacol Toxicol. 2006;98:320–3.
4. Moride Y, Haramburu F, Requejo AA, Bégaud B. Under-reporting of adverse drug reactions in general practice. Br J Clin Pharmacol. 1997;43:177–81.
5. Gaist D, Andersen M, Schou JS. Spontaneous reports of drug-induced erythema multiforme, Stevens–Johnson syndrome and toxic epidermal necrolysis in Denmark 1968–1991. Pharmacoepidemiol Drug Saf. 1996;5:79–86.
6. Alvarez-Requejo A, Carvajal A, Bégaud B, Moride Y, Vega T, Arias LH. Under-reporting of adverse drug reactions. Estimate based on a spontaneous reporting scheme and a sentinel system. Eur J Clin Pharmacol. 1998;54:483–8.
7. McCormick TH, Ferrell R, Karr AF, Ryan PB. Big data, big results: knowledge discovery in output from large-scale analytics. Stat Anal Data Min. 2014;7:404–12.
8. Hallas J, Pottegård A. Use of self-controlled designs in pharmacoepidemiology. J Intern Med. 2014;275:581–9.
9. Hallas J. Evidence of depression provoked by cardiovascular medication: a prescription sequence symmetry analysis. Epidemiology. 1996;7:478–84.
10. Wahab IA, Pratt NL, Wiese MD, Kalisch LM, Roughead EE. The validity of sequence symmetry analysis (SSA) for adverse drug reaction signal detection. Pharmacoepidemiol Drug Saf. 2013;22:496–502.
11. Wahab IA, Pratt NL, Ellett LK, Roughead EE. Sequence symmetry analysis as a signal detection tool for potential heart failure adverse events in an administrative claims database. Drug Saf. 2016;39:347–54.
12. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H, Sørensen HT, Hallas J, Schmidt M. Data resource profile: The Danish National Prescription Registry. Int J Epidemiol. 2016;46:798.
13. 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. Clin Epidemiol. 2015;7:449–90.
14. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. Eur J Epidemiol. 2014;29:541–9.
15. Thygesen LC, Daasnes C, Thaulow I, Brønnum-Hansen H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. Scand J Public Health. 2011;39:12–6.
16. Pratt NL, Ilomäki J, Raymond C, Roughead EE. The performance of sequence symmetry analysis as a tool for post-market surveillance of newly marketed medicines: a simulation study. BMC Med Res Methodol. 2014;14:66.
17. Tsiropoulos I, Andersen M, Hallas J. Adverse events with use of antiepileptic drugs: a prescription and event symmetry analysis. Pharmacoepidemiol Drug Saf. 2009;18:483–91.
18. Rothman KJ. Six persistent research misconceptions. J Gen Intern Med. 2014;29:1060–4.
19. Morris JA, Gardner MJ. Calculating confidence intervals for relative risks, odds ratios, and standardised ratios and rates. In: Gardner MJ, Altman DG, editors. Statistics with confidence. London: British Medical Journal Publishing; 1989. p. 60–1.
20. Cole DV, Kulldorff M, Baker M, et al. Infrastructure for evaluation of statistical alerts arising from vaccine safety data mining activities in mini-sentinel. https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_PRISM_Data-Mining-Infrastructure_Report_0.pdf.
21. Gruber S, Chakravarty A, Heckbert SR, Levenson M, Martin D, Nelson JC, et al. Design and analysis choices for safety surveillance evaluations need to be tuned to the specifics of the hypothesized drug-outcome association. Pharmacoepidemiol Drug Saf. 2016;25:973–81.
22. Bytzer P, Hallas J. Drug-induced symptoms of functional dyspepsia and nausea. A symmetry analysis of one million prescriptions. Aliment Pharmacol Ther. 2000;14:1479–84.
23. Lorberbaum T, Nasir M, Keiser MJ, Vilar S, Hripcsak G, Tatonetti NP. Systems pharmacology augments drug safety surveillance. Clin Pharmacol Ther. 2015;97:151–8.
24. Lai ECC, Pottegård A, Lin SJ, Hsieh C-Y, Hallas J, Yang Y-H. Antiepileptic drugs and risk of bacterial infections: a cross-national symmetry analysis from Denmark and Taiwan (submitted).
25. Harris DG. Management of pain in advanced disease. Br Med Bull. 2014;110:117–28.
26. Rasmussen L, Hallas J, Madsen KG, Pottegård A. Cardiovascular drugs and erectile dysfunction—a symmetry analysis. Br J Clin Pharmacol. 2015;80:1219–23.