Drug Safety (2020) 43:927–942
https://doi.org/10.1007/s40264-020-00950-3

ORIGINAL RESEARCH ARTICLE

Channeling Bias in the Analysis of Risk of Myocardial Infarction, Stroke, Gastrointestinal Bleeding, and Acute Renal Failure with the Use of Paracetamol Compared with Ibuprofen

Rachel B. Weinstein[1] · Patrick B. Ryan[1] · Jesse A. Berlin[2] · Martijn J. Schuemie[1] · Joel Swerdel[1] · Daniel Fife[1]

Published online: 4 June 2020

Abstract

Introduction Observational studies estimating severe outcomes for paracetamol versus ibuprofen use have acknowledged the specific challenge of channeling bias. A previous study relying on negative controls suggested that using large-scale propensity score (LSPS) matching may mitigate bias better than models using limited lists of covariates.

Objective The aim was to assess whether using LSPS matching would enable the evaluation of paracetamol, compared to ibuprofen, and increased risk of myocardial infarction, stroke, gastrointestinal (GI) bleeding, or acute renal failure.

Study design and setting In a new-user cohort study, we used two propensity score model strategies for confounder controls. One replicated the approach of controlling for a hand-picked list. The second used LSPSs based on all available covariates for matching. Positive and negative controls assessed residual confounding and calibrated confidence intervals. The data source was the Clinical Practices Research Datalink (CPRD).

Results A substantial proportion of negative controls were statistically significant after propensity score matching on the publication covariates, indicating considerable systematic error. LSPS adjustment was less biased, but residual error remained. The calibrated estimates resulted in very wide confidence intervals, indicating large uncertainty in effect estimates once residual error was incorporated.

Conclusions For paracetamol versus ibuprofen, when using LSPS methods in the CPRD, it is only possible to distinguish true effects if those effects are large (hazard ratio > 2). Due to their smaller hazard ratios, the outcomes under study cannot be differentiated from null effects (represented by negative controls) even if there were a true effect. Based on these data, we conclude that we are unable to determine whether paracetamol is associated with an increased risk of myocardial infarction, stroke, GI bleeding, and acute renal failure compared to ibuprofen, due to residual confounding.

1 Introduction

Numerous epidemiology studies over the past 3 decades have examined the risk of various serious adverse events such as renal failure, myocardial infarction (MI), stroke, and gastrointestinal (GI) bleeding among people exposed to paracetamol compared to ibuprofen and other over-the-counter (OTC) analgesics. The results have been inconsistent, and some have reported an increased risk [1–12].

Most prior epidemiologic studies discussed confounding, or channeling, as a possible source of bias influencing the results, but did not attempt to measure its impact. Channeling will direct "sicker" patients away from ibuprofen and toward paracetamol. Specifically, the label for ibuprofen notes GI bleed and heart and kidney disease, which would affect these outcomes in particular.

In a prior study, Weinstein et al. [13] examined the impact of channeling bias on 31 negative control outcomes, i.e., outcomes known not to be associated with paracetamol or ibuprofen use, where, therefore, the true hazard ratio (HR) is believed to be 1. Several models were used to show the impact of bias on the negative control associations with the implication that previous publications may have

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40264-020-00950-3) contains supplementary material, which is available to authorized users.

Rachel B. Weinstein
Rweinst1@its.jnj.com

[1] Janssen Research and Development, LLC, 1125 Harbourton-Trenton Rd, Titusville, NJ 08560, USA

[2] Johnson and Johnson, 1125 Harbourton-Trenton Rd, Titusville, NJ 08560, USA

Key Points

In comparative cohort studies to assess risk of myocardial infarction, stroke, gastrointestinal bleed, and renal failure in patients treated with paracetamol versus ibuprofen, results vary substantially depending on the models used to control for confounding and bias.

Large-scale propensity score matching resulted in attenuated effects and increased precision. However, substantial bias remained after large-scale propensity score matching, undermining the ability to discern or rule out an effect of exposure to paracetamol on these outcomes.

For comparisons of paracetamol versus ibuprofen, when using our methods in the Clinical Practices Research Datalink, it is only possible to distinguish true effects if those true effects are large (relative risk > 2).

inadequately adjusted for this bias. The results suggested that using large-scale propensity score (LSPS) matching may be a better way to reduce the impact of this bias than propensity score models based on a selected list of covariates. In the current study, we use the lessons from our earlier research to produce a more valid estimate than was provided by previous studies.

The primary objective of the current study was, therefore, to assess whether paracetamol, compared to ibuprofen, was associated with an increased risk of MI, stroke, GI bleeding, and acute renal failure. An additional objective was to assess the extent of residual bias in the estimation of the effect as specified in the primary objective using negative and positive controls [15, 16]. Any observed residual bias was to be incorporated into an empirically calibrated *p* value [14] and confidence interval (CI) [15].

2 Methods

In this study, we first replicated the findings seen in prior publications and then used a new adjustment strategy in a re-assessment. Both analyses used one-to-one propensity score matching. One propensity model was fit using variables from prior publications comparing paracetamol and ibuprofen outcomes. In a second propensity model, LSPS matching, with all available baseline covariates (over 10,000), was used and is referred to as the “full set of covariates available”. The full set of covariates characterized patient demographics, all prior conditions, drugs, procedures, and health service utilization patterns. The outcomes of interest were acute MI, stroke, GI bleed, and acute renal failure.

The analytic approach assessed on-treatment effects of the first treatment after a minimum 1-year period with no prescriptions for either paracetamol or ibuprofen. This allowed us to evaluate new prescriptions, which should capture any influence that baseline comorbidities might have on prescription of analgesics, i.e., channeling effects. Patients who were prescribed paracetamol alone and later, within the study period, prescribed ibuprofen were classified as paracetamol and then censored at the time of the ibuprofen prescription, and vice versa. Paracetamol users were one-to-one matched to ibuprofen users based on the propensity score. Cox proportional hazards models, conditioned on the matched sets, were used to compute the HRs for the outcomes of interest and the negative and positive controls [15, 16].

In this analysis, successful control of channeling by propensity score matching would be evidenced by the lack of statistically significant association of the negative control outcomes with the exposure, i.e., by observing no more negative controls associated with the exposure group than would be expected by chance, under the null hypothesis. The criteria for defining these negative controls are described below. Synthetic positive controls derived from the negative controls were used to estimate residual error at various effect sizes and to allow empirical calibration of CIs for the severe adverse event outcomes [15, 16]. The protocol and analysis source code for this study was posted on the Observational Health Data Sciences and Informatics (OHDSI) website Repository of OHDSI Collaborative Research Protocols (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ParacetamolvIbuprofen_MI).

2.1 Database Used

The data source for this analysis was the Clinical Practice Research Datalink (CPRD), a UK primary care database containing de-identified data from 1 January 1988 through 30 June 2017. There were 12.5 million people eligible during the timeframe to be included in the study prior to inclusion/exclusion criteria. The database includes data on demographics, conditions diagnosed, observations, measurements, hospitalizations, and procedures that the general practitioner (GP) is made aware of, in addition to any made by the GP. Prescriptions in the CPRD are not explicitly linked to their indications. A key strength of the data is the long-term follow-up, since overall, the median follow-up time for individual patients is about 5 years (interquartile range 1.8–11.1 years) [17]. Data for this study came from practices classified as “up-to-standard” (UTS) by the CPRD during the period of interest. The UTS designation reflects a minimum, practice-level measure of quality based on continuity of recording and number of deaths recorded. The protocol for this study (reference # 18_100R) was approved by the Independent Scientific Advisory Committee (ISAC).

The CPRD data in this analysis were converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) v5 [18]. The accuracy of Read codes, drugs, and clinical data was assessed, and a replication of a published case–control study confirmed the validity of the conversion [18]. Similar validations have been implemented in healthcare claims and hospital data [19, 20]. The full set of diagnosis and drug codes used in these analyses are available from the on-line protocol.

2.2 Cohort Definition

Patients were included in the study if they received a first prescription for either single-ingredient paracetamol or single-ingredient ibuprofen in 2005–2014. At the time the study was performed, this was the most recent 10-year period that allowed for up to 1 year of follow-up. Ten years of data were needed in order to adequately power the assessment in this on-treatment analysis. The date of this first prescription was defined as the index date. Patients were included if they were aged 18 or older on the index date and were enrolled for the 2 years prior to and at least 1 day after the index date. The 2-year observation period was intended to provide adequate information about medical history.

Reasons for a GP in the UK to prescribe a medication that is available OTC include the following: (1) to have a record of what was recommended to the patient; (2) to give the patient a reminder of what he/she is supposed to obtain and how it should be used; (3) to allow the patient to obtain the medication at a favorable price if, e.g., the patient needs a 30-day supply and a 30-day supply OTC costs more than the approximately US$10 that is the charge for filling a prescription; (4) to allow a patient to avail himself or herself of the right to have prescriptions filled free if the patient meets the income standard for doing so; and (5) to allow the patient to buy more than is permitted for OTC drugs.

To avoid observing prevalent prescription use, we required 12 months of continuous observation without prescriptions of single-ingredient paracetamol or ibuprofen prior to the index date. We classified analgesic use into two cohorts: (1) “paracetamol only”—patients with new, single-ingredient paracetamol exposure without concomitant ibuprofen, and (2) “ibuprofen only”—patients with new, single-ingredient ibuprofen exposure without concomitant paracetamol.

Patients who received prescriptions for both paracetamol and ibuprofen on their index date were excluded from the study. In addition, patients with a prescription for other nonsteroidal anti-inflammatory drugs (NSAIDs) or aspirin in the 12 months prior to or on the index date were excluded. Patients with a prescription for any paracetamol- or ibuprofen-containing combination products in the 12 months prior to and including the index date were excluded from the study. Patients with these concomitant and prior exposures were excluded to reduce confounding due to other products and to allow focus on the single ingredient.

2.3 Statistical Methods Used

2.3.1 Calculation of Time at Risk

Time at risk was calculated from the index date of first exposure to the end of treatment, based on days’ supply, allowing for up to 30-day gaps between the end of days’ supply and the start of the next prescription. The primary analysis was an on-treatment type analysis. In addition, two sensitivity analyses were performed by varying the treatment window, to understand the robustness of the results of the on-treatment requirement. The sensitivity analyses were analogous to intent-to-treat analyses with (1) a 90-day window and (2) a 1-year window from the start of exposure. For these analyses, exposure was assigned based on the drug at index and only censored if the patient received a prescription for the other drug, left the practice, or died before the end of the 90-day or 365-day window.

2.3.2 Outcomes of Interest

The four outcomes of interest were incident MI, stroke, GI bleeding, and acute renal failure. Variables for presence or absence of a given diagnosis were developed based on diagnoses in the database following the index date, regardless of the amount of time patients were in the database following the index date. We reviewed all available time prior to the outcome (including prior to the index exposure date) to determine whether an event was incident or not. People with a prior event were excluded from the analysis of that event only. We required all patients to have at least 1 day in the database after the index date. Disease codes were developed based on disease vocabularies available within the OMOP CDM, as well as clinician review, and can be found in the protocol posted on-line.

2.3.3 New-User Cohort Analysis Using Propensity Scores

A comparative cohort analysis was performed, comparing new users of paracetamol to new users of ibuprofen. Two propensity score models were fitted with different sets of variables. One was referred to as the “publication variables” and was intended to replicate analyses commonly seen in publications on adverse effects of paracetamol and/or ibuprofen [3, 4, 7–11]. The list of variables was extracted from a publication [3] that used the CPRD to study paracetamol and ibuprofen exposure. The publication variables propensity score model included the following: sex, age group (5-year increments), obese, morbidly obese, smoker, alcohol abuse, upper GI events, osteoarthritis, rheumatoid arthritis, ischemic heart disease, heart failure, hypertension, cerebrovascular disease, diabetes mellitus, hyperthyroidism, stroke or transient ischemic attack, cancer (excluding non-melanoma skin cancer), inflammatory bowel, autoimmune disease, depression, drug abuse, anticoagulants, oral glucocorticoids, diuretics, cardiac glycosides, statins, angiotensin receptor blockers, hypnotics, anxiolytics, antipsychotics, antibacterials, aminosalicylates, antidepressants, aspirin, oral corticosteroids, proton-pump inhibitors, histamine-2 receptor antagonists, hyperlipidemia, and NSAIDs (prior to the clean period).

In the second propensity score model, a larger set of baseline covariates was defined characterizing patient demographics, all prior conditions, drugs, procedures, and health service utilization patterns:

- *Demographics* Age (5-year increments), sex, index date year, and index date month
- *Conditions*[1] Presence/absence of condition in 365-day window prior to or on index date, presence/absence of condition in 30-day window prior to or on index date, presence/absence of condition diagnosed in inpatient stay in 180-day window prior to or on index date, presence/absence of an aggregation of episodes of care over time for a condition ("condition era") any time prior to or on index date (merging consecutive diagnosis codes into a single era, allowing for a maximum gap of 30 days between diagnoses), presence/absence of condition era overlapping the index date, presence/absence of an aggregation of episodes of care over time for a condition group [based on the systematized nomenclature of medicine (SNOMED) condition hierarchy] era any time prior to or on the index date, presence/absence of a condition group era overlapping the index date, Charlson Comorbidity Index score, Diabetes Complications Severity Index (DCSI) score, and CHADS2 score
- *Drugs*[2] Presence/absence of a length of time of exposure to a drug product ("drug era") in 365-day window prior to or on index date, presence/absence of drug era in 30-day window prior to or on index date, presence/absence of drug era overlapping the index date, presence/absence of drug era any time prior to or on index date, presence/absence of a drug group [using the Anatomical Therapeutic Chemical (ATC) hierarchy] era in a 30-day window prior to or on the index date, presence/absence of a drug group era overlapping the index date, and presence/absence of a drug group era any time prior to or on the index date
- *Observations*[3] Presence/absence of observation in 365-day window prior to or on index date, presence/absence of observation in 30-day window prior to or on index date, count of each observation concept in 365-day window prior to or on index date, and counts of the number of concepts a person has within each domain (i.e., condition, drug, procedure, clinical observation, laboratory measurement)
- *Procedures*[4] Presence/absence of procedure in 365-day window prior to or on index date and presence/absence of procedure in 30-day window prior to or on index date

We performed multivariable logistic regression to estimate a propensity score that predicted treatment (paracetamol vs. ibuprofen) using the publication variables defined above. Although the exact codes included in the definition of each variable were not detailed in prior publications, the authors indicated that the above conditions and exposures were controlled at baseline. The definitions of the publication variables for the current study are available from the online protocol (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ParacetamolvIbuprofen_MI).

For the LSPS-matched model, we used the large array of baseline covariates listed above to characterize patients at baseline. Because we have transformed the CPRD into the OMOP Common Data Model, v5, we applied standardized covariates within the open-source OHDSI application "CohortMethod" (https://ohdsi.github.io/CohortMethod/), which included SNOMED-coded concepts and higher-level classifications for conditions, and drugs coded at the RxNorm ingredient and ATC class levels. Baseline variables were evaluated based on data available prior to the cohort index date. We performed a regularized logistic regression to estimate a propensity score that predicted treatment assignment (paracetamol vs. ibuprofen). To avoid over-fitting and to accommodate the large number of predictors, an L1 penalty, i.e., least absolute shrinkage and selection operator (LASSO) [21, 22], was used. The optimal regularization hyper-parameter was estimated using tenfold cross-validation.

For the propensity-score models, we utilized the standardized mean differences (SMDs) for both the publication covariates and the full set of available covariates between paracetamol and ibuprofen in the two propensity score models as model diagnostics. SMDs measure covariate balance

[1] Examples of conditions based on the SNOMED hierarchy of conditions include essential hypertension, dyspnea, and osteoarthritis.

[2] Examples of drugs include analgesics, antithrombotic agents, and codeine. Covariates for both individual products as recorded in the database, as well as at the ingredient level. The products recorded in the CPRD include strength, form, and ingredients (e.g., 20-mg atorvastatin oral tablet), but not box size.

[3] An example of an observation is the tobacco smoking behavior.

[4] An example of a procedure is a radiologic examination.

between the two treatment groups and are the difference in prevalence in each cohort divided by the standard deviation. A large absolute value SMD on a covariate is an indication of significant disparity in the proportion of patients with the covariate between the two groups. An SMD > 0.1 is used as an ad-hoc heuristic for what constitutes "large" [23]. Preference scores, which are propensity scores normalized for imbalance in cohort size, were plotted [24]. Overlap of at least 50% of each cohort in the range 0.3–0.7 was taken to indicate clinical equipoise, i.e., patients are near the point of indifference in treatment selection.

The propensity scores were used to perform one-to-one matching (using a standardized caliper of 0.25 × propensity score standard deviation). The matched sets were used within a conditional (on matching) univariate Cox regression model, which estimated the effect of exposure on the incidence of each outcome, without further adjustment.

2.4 Negative Control Outcomes

Negative control outcomes are those determined a priori to have no causal relationship with the exposure of interest [25, 26]. We conducted an automated search of the medical literature and the relevant medication labels [27], followed by a clinical review. Thirty-nine outcomes believed to have no association with the exposures of interest were identified as negative controls. It is further assumed that confounding for the outcomes of interest is sampled from the same distribution as the confounding of the negative controls.

- Achilles tendinitis
- Atrophic vaginitis
- Breath smells unpleasant
- Bronchiectasis
- Disorders of initiating and maintaining sleep
- Ear problem
- Erythema nodosum
- Falls
- Foot-drop
- Ganglion and cyst of synovium, tendon and bursa
- Hemangioma
- Hydrocele
- Hyperthyroidism
- Impaired glucose tolerance
- Impingement syndrome of shoulder region
- Impotence
- Incontinence of feces
- Interpersonal relationship finding
- Irregular periods
- Irritability and anger
- Joint stiffness
- Loss of sense of smell
- Mixed hyperlipidemia
- Osteitis deformans
- Panic attack
- Perforation of tympanic membrane
- Pes planus
- Polymyalgia rheumatica
- Premature menopause
- Prolapse of female genital organs
- Pure hypercholesterolemia
- Respiratory symptom (only inclusive of the concepts respiratory symptom, snoring symptoms, and complaint of postnasal drip)
- Restless legs
- Restlessness and agitation
- Rosacea
- Simple goiter
- Skin sensation disturbance
- Snapping thumb syndrome
- Urinary symptoms

Models that adequately control for confounding factors should produce HR estimates of the null value (1.0) for these negative control outcomes. These models allow for the examination of the extent of bias in the data which is reflected in the degree that they produce significant HR estimates different from 1.0. Under the null hypothesis, with little to no residual bias, we would expect no more than one or two HRs to be significantly different from unity based on a p value of less than 0.05. An empirical distribution of the HR under the null was developed.

2.5 Positive Control Outcomes

In addition to negative control outcomes, we also included synthetic positive control outcomes. Positive control outcomes were based on the negative controls described above, but the true effect size was artificially increased to a desired effect size by injection of additional, simulated outcomes. To preserve confounding, these additional outcomes were sampled from predicted probabilities generated using a fitted predictive model. For each negative control outcome, three positive control outcomes were generated with true RRs of 1.5, 2, and 4. The residual error estimated from these positive controls was used to perform CI calibration [15, 16].

3 Results

The total number of patients in the CPRD dataset eligible for inclusion in the study between January 1, 2005 and December 31, 2014 was 9,365,462. The number of eligible patients during this time period, with new exposures for paracetamol and ibuprofen, was 288,967 and 382,749, respectively. The loss of patients differed slightly for each incident event analysis due to the exclusion of those with a prior event (Fig. 1; Table 1). Following propensity score matching on the publication covariates, 56% of the original sample population remained in the study. Following propensity score matching on the full set of covariates, 40% of the original sample population remained in the study.

The distributions of preference scores for paracetamol and ibuprofen using the publication covariates are shown in Fig. 2a. The overlap in the preference scores between 0.3 and 0.7 was greater than 50% for the publication covariates (Table 2). Matching on this set of covariates would, technically, fulfill the criteria for clinical equipoise in a comparative effectiveness assessment for the two treatment options. However, when utilizing the full set of covariates (Fig. 2b),

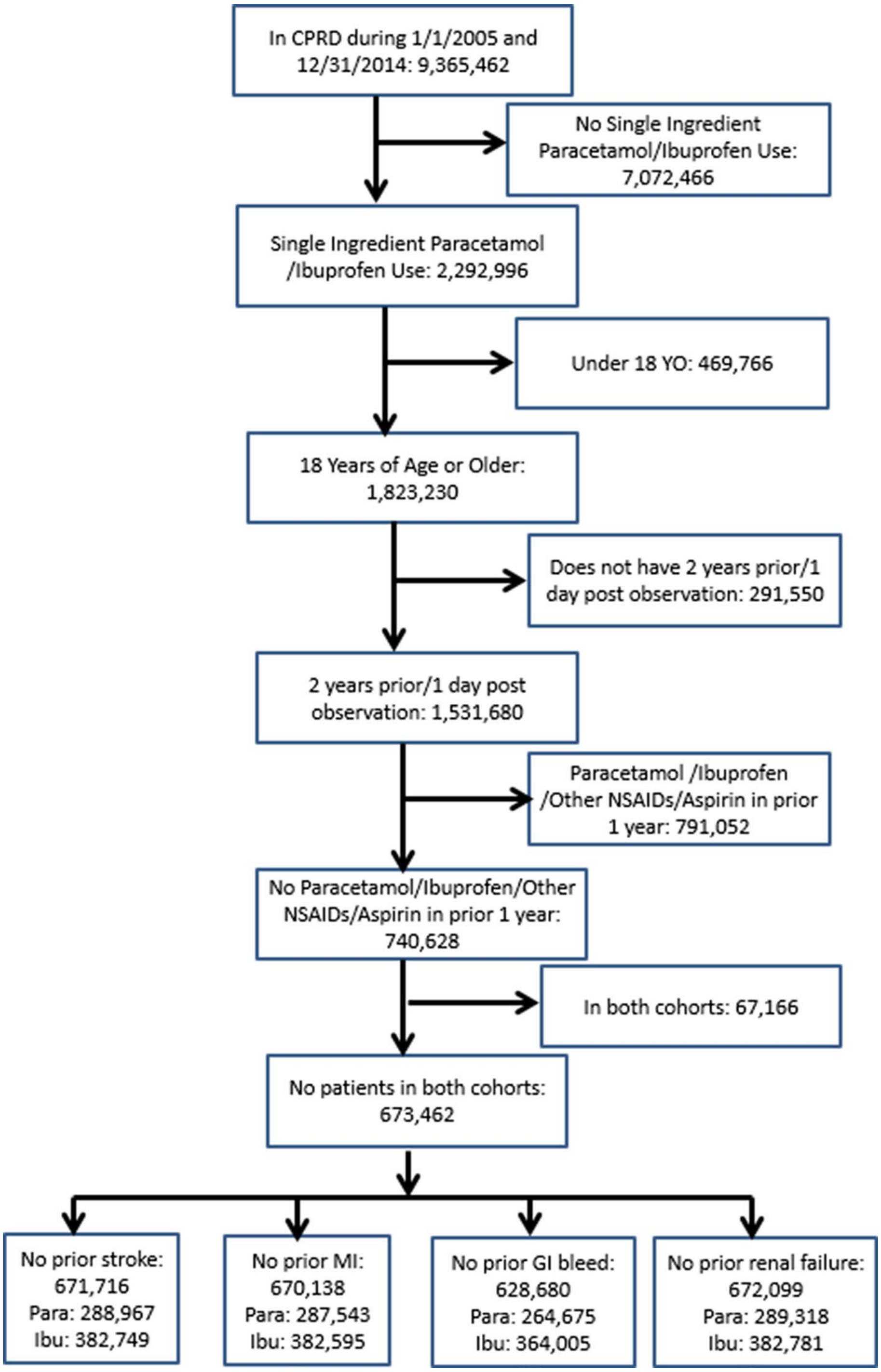
In CPRD during 1/1/2005 and 12/31/2014: 9,365,462
No Single Ingredient Paracetamol/Ibuprofen Use: 7,072,466
Single Ingredient Paracetamol /Ibuprofen Use: 2,292,996
Under 18 YO: 469,766
18 Years of Age or Older: 1,823,230
Does not have 2 years prior/1 day post observation: 291,550
2 years prior/1 day post observation: 1,531,680
Paracetamol /Ibuprofen /Other NSAIDs/Aspirin in prior 1 year: 791,052
No Paracetamol/Ibuprofen/Other NSAIDs/Aspirin in prior 1 year: 740,628
In both cohorts: 67,166
No patients in both cohorts: 673,462
No prior stroke: 671,716 Para: 288,967 Ibu: 382,749
No prior MI: 670,138 Para: 287,543 Ibu: 382,595
No prior GI bleed: 628,680 Para: 264,675 Ibu: 364,005
No prior renal failure: 672,099 Para: 289,318 Ibu: 382,781

Fig. 1 Flow of patients from the Clinical Practice Research Datalink (CPRD) to the analytic study population. *GI* gastrointestinal, *Ibu* ibuprofen, *MI* myocardial infarction, *NSAID* nonsteroidal anti-inflammatory drug, *Para* paracetamol *YO* years old

Table 1 Sample sizes for each outcome of interest before and after propensity score matching

	Stroke		MI		GI bleed		Renal failure	
	N	% Matched	*N*	% Matched	*N*	% Matched	*N*	% Matched
Before matching	671,716		670,138		628,680		672,099	
Publication covariates matched	374,274	55.7	374,022	55.8	349,212	55.5	374,336	55.7
Full set of covariates matched	265,820	39.6	265,550	39.6	249,348	39.7	265,912	39.6

GI gastrointestinal, *MI* myocardial infarction

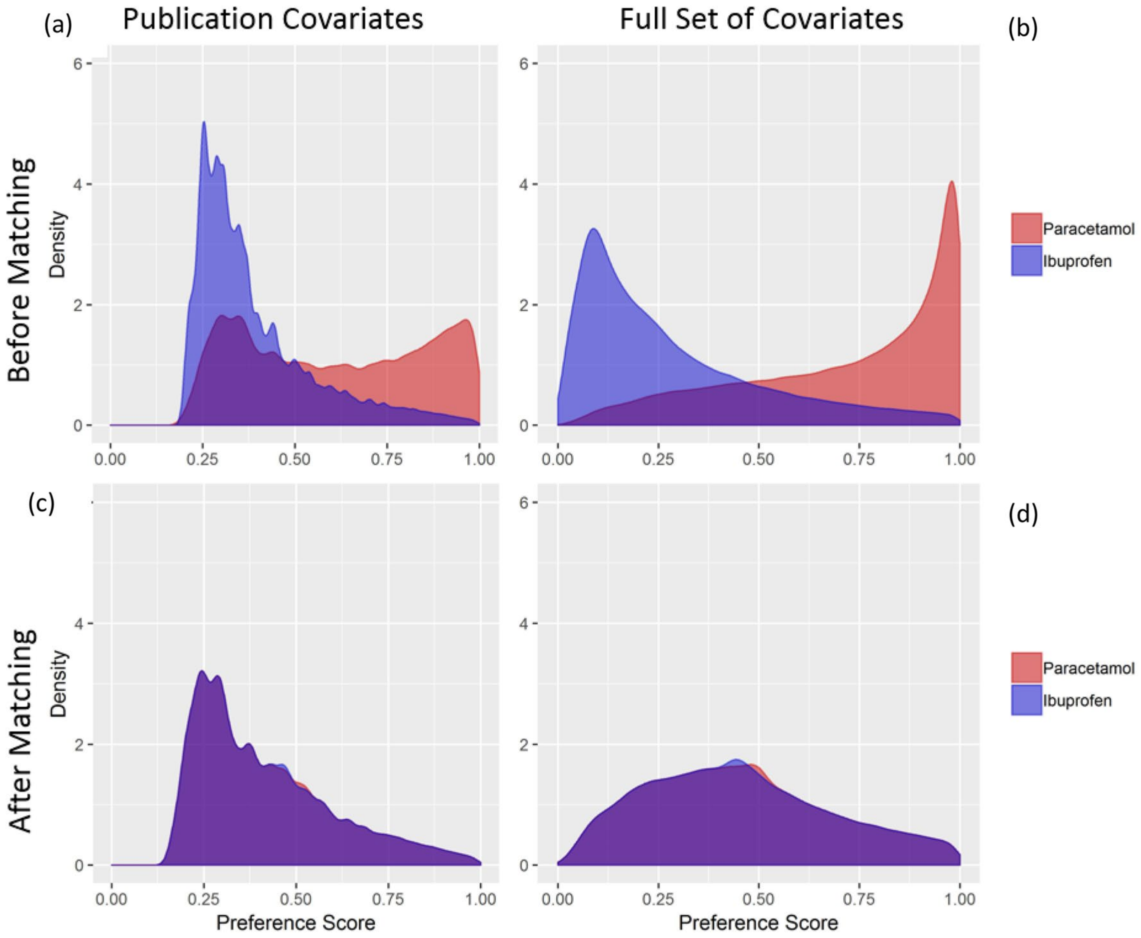

Fig. 2 Distribution of propensity scores **a** and **b** before and **c** and **d** after matching from publication covariates and the full set of covariates for paracetamol and ibuprofen cohorts

the proportion of overlap dropped below 50% and was evidence that the majority of patients were not near clinical equipoise with respect to the selection of paracetamol versus ibuprofen.

Selected demographic, prior condition, and drug characteristics of the pre- and post-matching populations are shown in Table 3. Prior to matching, there was a large difference in age distributions between the paracetamol (mean age 61 years) and ibuprofen (mean age 46 years) groups, which was reduced after matching (mean age approximately 62 across all groups). The proportion with each diagnosis and

Table 2 Propensity score using publication and full set of covariates with preference score distribution between 0.3 and 0.7

Analgesic	All scores	0.3 ≤ preference ≤ 0.7	
	N	*N*	%
Publication covariates			
Paracetamol	288,967	138,273	47.9
Ibuprofen	382,749	220,442	57.6
Full set of covariates			
Paracetamol	288,967	85,800	29.7
Ibuprofen	382,749	107,453	28.1

Table 3 Distribution of selected characteristics 1 year prior to first use of paracetamol and ibuprofen during 2005–2015 among users in the study populations before and after matching on the publication variables propensity score model and the large-scale propensity score model

Characteristic	Before matching			After matching on publication covariates			After matching on full set of covariates		
	Paracetamol	Ibuprofen	Std. diff	Paracetamol	Ibuprofen	Std. diff	Paracetamol	Ibuprofen	Std. diff
Number	288,967	382,749		187,137	187,137		132,910	132,910	
Female %	62.9	57.8	0.105	61.5	62.5	− 0.020	61.0	61.7	− 0.014
Mean age, years (SD)	61.2 (20.6)	46.3 (17.3)		52.3 (18.9)	52.0 (18.9)		52.7 (19.5)	52.9 (19.3)	
Age, years %									
15–19	2.6	5.1	−0.133	4.0	4.1	−0.004	4.4	4.4	−0.001
20–24	4.2	7.7	−0.148	6.0	6.1	−0.005	6.1	5.9	0.005
25–29	4.3	7.4	−0.132	6.1	6.3	−0.009	5.8	5.6	0.007
30–34	4.4	7.9	−0.144	6.4	6.6	−0.009	6.1	6.0	0.002
35–39	4.5	9.0	−0.181	6.4	6.4	−0.002	6.4	6.2	0.009
40–44	4.7	9.8	−0.200	6.6	6.5	0.003	6.8	6.5	0.010
45–49	4.9	9.4	−0.175	6.9	6.7	0.006	6.9	6.7	0.008
50–54	5.3	8.4	−0.125	7.1	7.0	0.007	7.0	6.8	0.006
55–59	6.0	7.7	−0.070	7.8	7.7	0.004	7.3	7.4	−0.003
60–64	9.4	9.7	−0.009	11.8	11.9	−0.004	11.2	11.7	−0.013
65–69	9.5	6.8	0.098	10.5	10.6	−0.004	9.9	10.4	−0.017
70–74	10.0	4.8	0.201	8.8	8.7	0.001	8.4	8.7	−0.012
75–79	10.2	3.2	0.282	6.1	5.9	0.005	6.5	6.7	−0.008
80–84	9.4	1.9	0.331	3.4	3.3	0.003	4.2	4.2	0.004
85–89	6.6	0.8	0.307	1.5	1.4	0.005	2.1	1.9	0.010
90–94	3.1	0.3	0.222	0.5	0.5	0.006	0.7	0.6	0.011
95–99	0.9	0.1	0.120	0.1	0.1	0.005	0.2	0.1	0.009
Index year (%)									
2005	12.7	14.1	−0.041	12.3	14.0	−0.052	13.4	13.3	0.005
2006	11.2	13.0	−0.054	11.0	12.9	−0.059	12.1	11.8	0.010
2007	10.2	11.8	−0.050	9.9	11.7	−0.056	10.7	10.4	0.009
2008	10.3	11.2	−0.030	10.2	11.0	−0.026	10.5	10.6	−0.004
2009	10.1	10.7	−0.021	10.2	10.6	−0.015	10.2	10.1	0.004
2010	9.9	10.0	−0.003	10.1	9.9	0.008	9.8	9.8	0.000
2011	9.5	9.2	0.010	9.5	9.1	0.015	9.3	9.2	0.002
2012	9.8	8.0	0.060	10.0	8.2	0.062	9.2	9.5	−0.012
2013	8.9	6.7	0.081	9.1	6.9	0.079	8.1	8.3	−0.006
2014	7.6	5.4	0.088	7.6	5.6	0.082	6.8	7.1	−0.011
Conditions, %									
Smoker	2.0	2.0	0.000	2.2	2.2	0.005	2.2	2.1	0.001
Osteoarthritis	19.4	9.4	0.286	14.5	13.4	0.030	14.4	14.5	−0.004
Neoplasm of prostate	1.1	0.4	0.078	0.6	0.7	−0.016	0.7	0.7	−0.002
Hyperthyroidism	0.9	0.6	0.031	0.7	0.7	0.002	0.7	0.7	0.001
Inflammatory bowel disease	0.6	0.3	0.055	0.5	0.3	0.019	0.4	0.4	−0.002
Rheumatoid arthritis	1.0	0.4	0.077	0.8	0.6	0.024	0.6	0.7	−0.006
Heart failure	2.3	0.2	0.192	0.5	0.3	0.028	0.5	0.4	0.011
Ischemic heart disease	3.9	0.6	0.220	1.3	1.0	0.021	1.3	1.2	0.011
Essential hypertension	16.1	7.9	0.254	11.5	11.8	−0.009	11.7	12.1	−0.014
Hyperlipidemia	7.3	4.1	0.141	5.7	5.7	−0.001	5.7	5.8	−0.005
Cerebrovascular disease	3.2	0.4	0.209	0.8	0.7	0.015	0.9	0.8	0.005
Drugs, %									
Cardiac glycosides	3.0	0.2	0.229	0.4	0.3	0.020	0.5	0.4	0.020

Table 3 (continued)

Characteristic	Before matching			After matching on publication covariates			After matching on full set of covariates		
	Paracetamol	Ibuprofen	Std. diff	Paracetamol	Ibuprofen	Std. diff	Paracetamol	Ibuprofen	Std. diff
Corticosteroids for systemic use	17.2	10.2	0.205	13.3	13.6	−0.011	13.0	13.1	−0.003
Antibacterials for systemic use	51.7	41.7	0.201	48.4	49.0	−0.011	47.2	47.7	−0.010
Proton pump inhibitors	23.8	14.1	0.250	19.5	19.6	−0.001	17.8	18.2	−0.010
Agents acting on the renin-angiotensin system	21.5	8.6	0.365	13.5	13.5	−0.002	13.7	13.9	−0.008
Other NSAIDs and nonsteroidal anti-rheumatics[a]	1.2	1.1	0.015	1.3	1.2	0.008	1.3	1.3	−0.001
Other analgesics[a]	16.5	5.8	0.344	16.0	6.7	0.296	10.0	9.9	0.003
Anxiolytics	7.9	5.8	0.085	7.0	7.2	−0.008	6.7	6.6	0.001
Hypnotics and sedatives	9.3	4.8	0.176	7.2	6.6	0.026	6.7	6.7	0.002
HMG-CoA reductase inhibitors	18.8	8.1	0.316	12.4	12.5	−0.001	12.6	12.8	−0.006
Vitamin K antagonists	7.2	0.5	0.356	1.4	0.9	0.056	1.4	1.1	0.028

NSAID nonsteroidal anti-inflammatory drug, *SD* standard deviation, *Std. diff* standardized difference

[a]These were from the period prior to 12 months prior to index date

drug was higher in the paracetamol cohort than the ibuprofen cohort, reflecting channeling, which was reduced after matching. In general, matching on either the publication or the full set of covariates reduced the differences between the cohorts.

The reductions in the SMDs for the covariates between cohorts following matching varied (Fig. 3). Prior to matching on the publication covariates (Fig. 3a), the SMDs ranged up to about 0.42 (*x*-axis), indicating appreciable differences in the distributions of these covariates between the two populations. Following matching, the SMDs ranged to less than 0.05 (*y*-axis), indicating more balance in baseline characteristics between the populations.

As an indication of the residual difference in the data after matching on publication variables, the SMDs for the full set of covariates in the model using propensity scores developed only from the publication covariates (Fig. 3b) shows poor balance both before and after matching. Prior to matching, the SMDs ranged up to about 0.5. After matching the SMDs were still as high as 0.35, indicating a substantial imbalance among variables not represented in the past publications, despite adequate balance of the publication variables. The following are representative of the covariates for which the

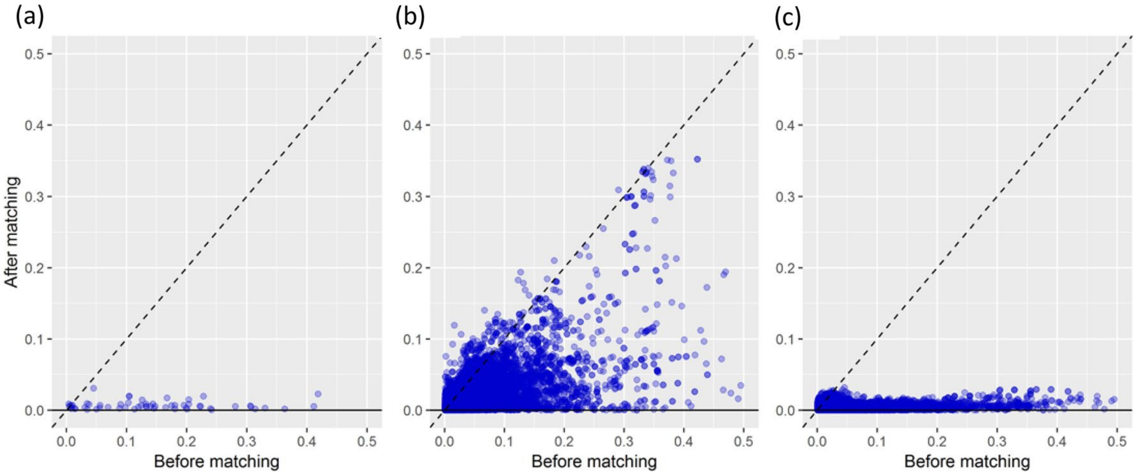

Fig. 3 Scatter plot of the covariate balance standardized mean difference **a** publication covariates before and after matching on the publication variable model, **b** covariate balance of all covariates before and after matching on the publications variable model, and **c** covariate balance of all covariates before and after matching on the all covariates model

SMD was among the highest: pain at a specific anatomical site, opioid prescription on or in the 30 days prior to the index date, prescription for cough and cold preparations on or prior to the index date, and prescription for codeine on or prior to the index date.

For the full set of covariates (Fig. 3c), prior to matching, the SMDs varied up to about 0.5 in the publication covariates. Following matching, the SMDs were no more than 0.05.

The area under the receiver operating characteristics curve (ROC AUC) was derived from the models limited to the publication covariates and, separately, from the models including the full set of covariates. The full set of covariates model provided better discrimination (AUC = 0.88) between the patients receiving paracetamol and those receiving ibuprofen compared to the publication covariates model (AUC = 0.77).

The distribution of exposure time for patients exposed to either paracetamol or ibuprofen was examined (On-line Fig. 1; see the electronic supplementary material). The mean ages of the two cohorts differed substantially. Over 75% of each cohort was exposed for 30 days or fewer. By 90 days after the index date, less than 10% of each cohort continued to be exposed to either paracetamol or ibuprofen.

To check that the proportional hazards assumption held for each of the analyses, Kaplan–Meier plots for the on-treatment exposure for each outcome of interest were generated (On-line Fig. 2; see the electronic supplementary material). The curves suggest that the proportional hazards assumption holds for analyses of stroke, MI, and GI bleeding, but may not hold for renal insufficiency.

The results of the primary analysis are shown in Table 4 and Fig. 4. After matching on the publication covariates propensity score model, all outcomes resulted in a statistically significant uncalibrated HR of paracetamol versus ibuprofen, except for MI (HR 1.48, 95% CI 0.88–2.54). For stroke, the HR was 2.67 (95% CI 1.10–7.43), for GI bleed it was 1.81 (95% CI 1.49–2.20), and for renal failure it was 4.86 (95% CI 2.29–11.95). In contrast, in the models after matching using all available covariates in the propensity score model, the HRs were attenuated compared to those based on the publication variables and none of the outcomes were statistically significant using traditional (uncalibrated) p value calculations, with the exception of GI bleed (HR 1.36, 95% CI 1.10–1.68). The HR for stroke was 2.20 (95% CI 0.80–6.98), for MI was 1.29 (95% CI 0.69–2.47) and for renal failure was 1.25 (95% CI 0.59–2.72). Using calibrated p values and CIs, none of the associations were statistically significant (see Table 4).

Figure 4 shows scatter plots of the HR by standard error from the Cox proportional hazard models with calibrated confidence regions for both adjustment approaches. The figures include the negative control HR outcomes (blue points) and the HRs for the outcomes of interest (yellow diamonds). Estimates below the dashed line (gray area) have $p < 0.05$ using the traditional p value calculation. Estimates below the dark orange line have $p < 0.05$ using the calibrated p values. The shaded region around the dark orange line represents the uncertainty (as reflected by a 95% credible interval) in the empirical calibration. The extent of systematic error is demonstrated in the distance the negative controls are from the vertical line at 1.0 and the number that fall in the traditional significance region (below the dashed line). The asymmetry of the distribution around 1.0 is evidence of bias.

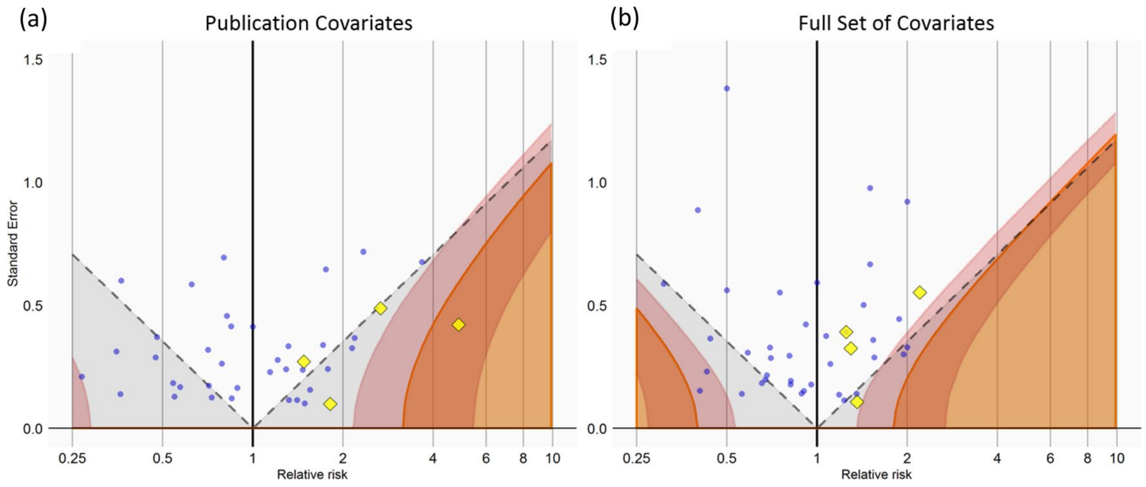

Fig. 4 Hazard ratio and standard error estimates from a propensity score matched comparison on-treatment analysis between paracetamol and ibuprofen users using either **a** publication covariates or **b** the full set of covariates. Estimates below the dashed line (grey area) have a p value of < 0.05 using traditional p value calculation. Estimates below the *orange line* have a p value of < 0.05 using the calibrated p values. Blue dots indicate negative controls, and yellow diamonds indicate the outcomes of interest

Table 4 Cox proportional hazard analysis of on-treatment exposure to paracetamol vs. ibuprofen on stroke, MI, GI bleeding, and renal failure (non-chronic) with patients matched by either publication covariates or the full set of covariates

	Propensity score matched		Calibrated estimates	
	HR (95% CI)	Null *p*	HR (95% CI)	Null *p*
Publication covariates				
Stroke	2.67 (1.10–7.43)	**0.045**	3.53 (0.51–62.91)	0.154
MI	1.48 (0.88–2.54)	0.150	1.89 (0.28–21.78)	0.405
GI bleed	1.81 (1.49–2.20)	**< 0.001**	2.34 (0.42–27.27)	0.241
Renal failure	4.86 (2.29–11.95)	**< 0.001**	6.65 (1.17–143.67)	**0.026**
Full set of covariates				
Stroke	2.20 (0.80–6.98)	0.154	2.89 (0.60–35.84)	0.155
MI	1.29 (0.69–2.47)	0.428	1.60 (0.41–10.27)	0.397
GI bleed	1.36 (1.10–1.68)	**0.004**	1.68 (0.60–9.12)	0.232
Renal failure	1.25 (0.59–2.72)	0.569	1.54 (0.34–10.62)	0.475

Bold text indicates statistically significant results

CI confidence interval, *GI* gastrointestinal, *HR* hazard ratio, *MI* myocardial infarction

The negative controls showed there was considerable confounding in the data (Fig. 4) and were observed to produce estimates anywhere in the range from HR = 0.5 to 2 that would meet traditional conventions of statistical significance at $p < 0.05$. The rate of error was well beyond the 5% false positive expected by chance alone. The residual confounding under the publication variables model (Fig. 4a) was substantial. The residual confounding under the full set of covariates propensity score model (Fig. 4b) was also substantial, though less so than for the publication covariates model. The graphs in Fig. 4 also show the outcomes of interest (yellow diamonds), which were mostly not significant after propensity score matching according to the traditional *p* value cutoff (except for renal failure in the publications model) and after calibration (exceed the dark orange line).

We conducted two sensitivity analyses assessing both 90-day and 365-day paracetamol and ibuprofen exposures. These results are shown in Table 5 and Figs. 5 and 6. A similar pattern was seen here where the models using full covariate matching had consistently smaller effect estimates relative to the publication covariates for all outcomes. The CIs in the LSPS-based models were narrowed compared to the publication covariate models, showing greater precision. However, for the calibrated estimates, the negative controls indicated that there was still substantial risk of residual confounding, preventing valid statistical interpretation of the CIs.

4 Discussion

In this study, we assessed whether using LSPS matching would enable estimation of the causal association of paracetamol versus ibuprofen for risk of MI, stroke, GI bleed, and acute renal failure. In a prior study of paracetamol and ibuprofen [13], the objective of which was to assess channeling bias, LSPS matching was better at controlling for residual confounding in estimating associations between exposure and negative control outcomes than the use of a selected list of variables in the propensity score model. As in the prior study, this study showed that using LSPS matching improved control for confounding, as evidenced by examination of the covariate balance. Additional confounding adjustment using the LSPS model had a material impact by reducing the effect estimates for these outcomes.

Matching on publication covariates and using on-treatment time at risk in the Cox models showed much larger residual bias than matching on LSPS. The change in significance of the outcomes of interest between the two Cox models (publication covariates vs. full set of covariates) suggests that the more complete approach to confounding control substantially impacts study results and raises questions about conclusions from earlier comparative outcome studies that used selected covariates for confounder control. It is important to point out that we capture covariates before (and on) the day of treatment assignment. Therefore, even if the covariates are differentially misclassified, one would expect decreases in precision (i.e., power), not increases in bias.

Even after matching on LSPS, the HRs for the negative controls remained very broadly dispersed and over 20% fell into the traditional significance region, violating the premise of a 5% probability of significance by chance alone under the null hypothesis and indicating substantial residual bias. After *p* value calibration, no outcomes of interest were statistically significant. Given this dispersion of the negative controls, we can neither confirm nor reject the possibility of an increased risk of these outcomes with paracetamol compared to ibuprofen. This does not imply there is no effect, just that given the observed residual bias, using these methods on these data, we are not able to discern an effect. In other

Table 5 Cox proportional hazard analysis of 90-day or 365-day intent to treat exposures to paracetamol vs. ibuprofen on stroke, MI, GI bleeding, and renal failure (non-chronic) with patients matched by either publication covariates or the full set of covariates

	90-day intent to treat			365-day intent to treat		
	HR (95% CI)	Null *p*	Calibrated *p*	HR (95% CI)	Null *p*	Calibrated *p*
Publication covariate matched						
Stroke	2.05 (1.22–3.57)	**0.009**	0.195	1.78 (1.31–2.43)	**< 0.001**	0.104
MI	1.12 (0.85–1.49)	0.432	0.699	1.15 (0.99–1.33)	0.071	0.578
GI bleed	1.44 (1.28–1.63)	**< 0.001**	0.409	1.31 (1.22–1.40)	**< 0.001**	0.358
Renal failure	3.12 (2.03–4.94)	**< 0.001**	**0.045**	2.58 (2.00–3.35)	**< 0.001**	**0.009**
Full set of covariates matched						
Stroke	1.65 (0.91–3.07)	0.107	0.173	1.15 (0.81–1.64)	0.426	0.441
MI	1.12 (0.82–1.54)	0.474	0.559	1.00 (0.84–1.19)	1.000	0.803
GI bleed	1.12 (0.97–1.28)	0.114	0.524	1.14 (1.05–1.24)	**0.001**	0.298
Renal failure	1.35 (0.81–2.26)	0.253	0.334	1.54 (1.14–2.10)	**0.005**	**0.036**

Bold text indicates statistically significant results

CI confidence interval, *GI* gastrointestinal, *HR* hazard ratio, *MI* myocardial infarction

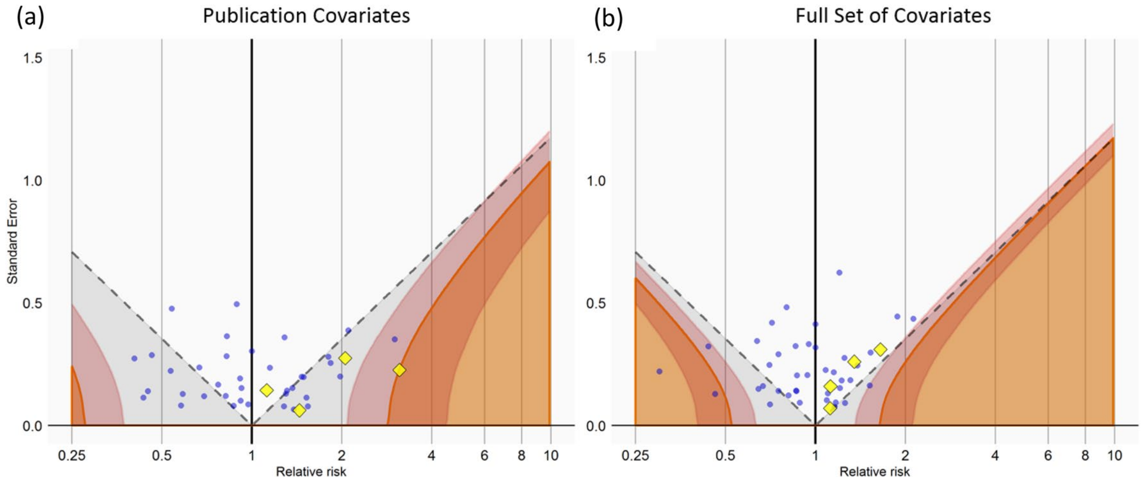

Fig. 5 Hazard ratio and standard error estimates from a propensity score matched comparison of 90 day exposure analyses between paracetamol and ibuprofen users using either **a** publication covariates or **b** the full set of covariates

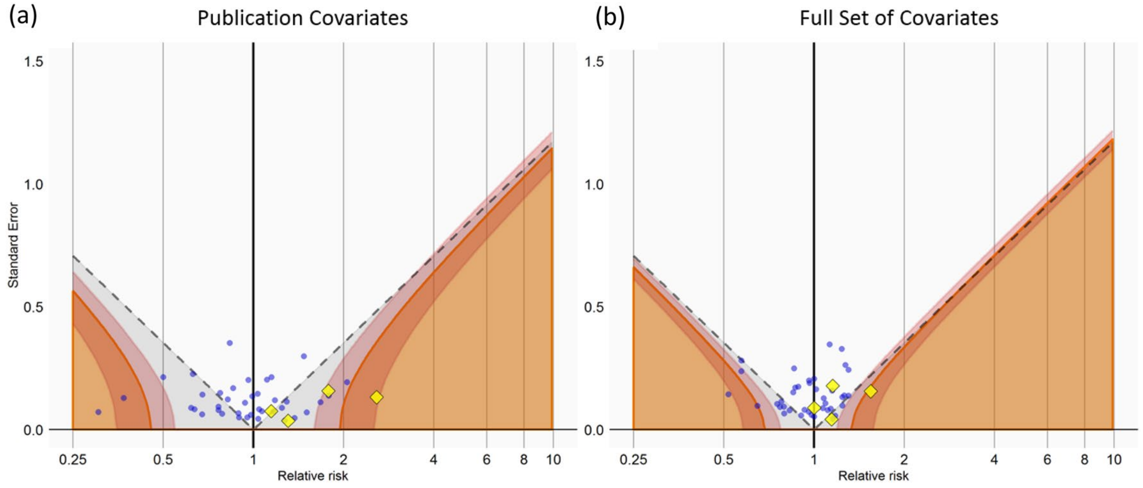

Fig. 6 Hazard ratio and standard error estimates from a propensity score matched comparison of 365 day exposure analyses between paracetamol and ibuprofen users using either **a** publication covariates or **b** the full set of covariates

studies [28–30], where we used negative controls selected in a similar way, we observed far less residual bias, suggesting that it is the comparison between paracetamol and ibuprofen that is problematic; neither of the two approaches to adjust for confounding were able to fully adjust for the inherent differences between these two exposure groups.

The purpose of the sensitivity analyses was to explore the robustness of the on-treatment results by repeating the analyses under 90-day and 365-day risk windows. Under these scenarios, the observation windows were increased beyond the prescription duration, which was 30 days, on average. With the expanded time-at-risk window, the number of events increased, leading to narrower CIs; however, the dispersion of the negative controls, while reduced, demonstrated substantial systematic error in comparison with the outcomes of interest. In the end, it was not possible to discriminate RRs of the outcomes of interest from the negative controls.

There have been numerous prior observational cohort studies examining adverse outcomes in paracetamol use with inconsistent results [1–3, 7–11]. The extent to which they attempted to control for confounding varied, though all were likely influenced by residual confounding. A few examples are offered here. One was a Danish population-based cohort study of cause of death among adult paracetamol users [2]. They used a standardized mortality ratio (SMR) analysis, which only controlled for age group, to compare to the general population. Mortality was significantly elevated for all causes of death examined, including specific cancers (e.g., breast, prostate, ovary), liver and renal disease, and cardiovascular diseases. SMRs were highest in the first year after prescription and tended toward 1.0 with increasing years of follow-up. Patterns by number of prescriptions were essentially flat for most outcomes except renal failure (a known adverse event for NSAIDs), which increased risk with more prescriptions. The authors noted that cancer is a “registered indication for use of paracetamol in Denmark” and also that standard textbooks in Denmark recommend paracetamol for first-line treatment of pain in chronic, nonmalignant conditions, and cancer, in office practice as well as in hospital. According to the authors, these patterns of association were likely due to confounding with the indication, and stated future studies would need to evaluate such confounding in the assessment of causal association.

Several studies have examined incident hypertension and change in kidney function [creatinine and estimated glomerular filtration rate (eGFR)] in established cohorts such as the Nurses’ Health Study (NHS) and the Physicians’ Health study (PHS) [7, 9–11]. Curhan et al. [11] used the NHS to examine the effects of lifetime use of paracetamol, aspirin, and NSAIDs in an 11-year follow-up study, adjusting for selected covariates at baseline. They reported a significant decline in GFR with increased lifetime paracetamol use. However, no significant differences were found for aspirin and NSAIDs. Given the warnings about renal disease in the labels for NSAIDS, this pattern is suggestive of channeling and confounding by contraindication.

In another NHS study, Dedier et al. [9] examined the risk of incident hypertension with the use of paracetamol, aspirin, and NSAIDs, compared to no use (of each analgesic in separate regressions) over an 8-year follow-up period. After adjusting for selected covariates, the odds ratios (ORs) for incident hypertension among women with the highest frequency of use (22 days/month) were significantly increased for paracetamol, aspirin, and NSAIDs. For each analgesic type, there was a significant trend in frequency of use.

Chan et al. [4] using the NHS examined frequency of use of paracetamol, NSAIDs, and aspirin on the risk of cardiovascular events including nonfatal MI, nonfatal stroke, fatal coronary heart disease, and fatal stroke. In multivariate models with selected covariates, including other analgesics, compared to no use, using at least 22 days per month the relative risk (RR) for paracetamol was 1.35 (95% CI 1.14–1.59). Frequent (≥ 22 days/month) use of NSAIDs compared to no use was also associated with increased risk for cardiovascular events: RR = 1.44 (95% CI 1.27–1.65). The results for aspirin were not significant and did not suggest increased risk for cardiovascular outcomes.

Kurth et al. [7] used the PHS to examine creatinine and GFR associated with analgesics use (lifetime total number of pills) compared to no use in a 14-year follow-up study. After adjusting for selected baseline covariates including total analgesic use (one of the few that did), ORs for paracetamol, aspirin, and other NSAIDs were not significantly associated with increased creatinine levels or decreased GFRs.

De Vries et al. [3] used the same data source as the present study to examine risk of several outcomes, including MI, stroke, GI bleeding, renal failure, congestive heart failure, and mortality, for prescription of paracetamol alone, ibuprofen alone, and concomitant exposure (both paracetamol and ibuprofen prescribed on the same day). Recognizing the presence of substantial heterogeneity, they tested the robustness of the data by analyzing risks for varying dosage patterns (first prescription, long gap, medication possession ratio) in current use. For MI, stroke, and renal failure, the RR for first prescription of current versus past use was statistically significant for paracetamol alone (without concomitant ibuprofen), in multivariable models adjusted for selected covariates. The corresponding RRs for ibuprofen alone and concomitant with paracetamol were only significant for renal failure.

Most recently, Roberts et al. [1] conducted a systematic review of observational studies of adverse effects of paracetamol compared with non-use. Because exposure and outcome measures differed across studies, combined estimates could only be generated (online supplement) for incident hypertension based on Curhan et al. [10] and Dedier et al. [9] and showed significant results for frequencies over 5 days per month. Forest plots for other outcomes by dose, frequency, or quantity suggested associations. While acknowledging that channeling bias may have played an important role, they did not consider that bias due to severity of disease or symptoms being treated were also consistent with the dose–response results.

The sensitivity analyses to understand the robustness of results included a 90-day and 365-day treatment window. The 90-day risk window is plausible for medications taken as needed, while a 365-day risk window is less plausible. The sensitivity analyses risked introducing bias by lengthening the time between measurement of the exposure, baseline covariates, and outcome and increasing the possibility that confounders were misclassified. These sensitivity analyses may more closely replicate findings from prior cohort studies with long follow-up times [3, 4, 7, 9–11] than the on-treatment analyses. Curhan et al. [11] examined renal function decline over a period of 11 years. Chan et al. [4] followed participants for 12 years, and Dedier et al. [9] followed participants for 8 years. Kurth et al. [7] followed participants in the PHS for 14 years. Lipworth et al. [2], using a retrospective database, had a mean follow-up of 3.5 years since first prescription. De Vries et al. [3] had a mean follow-up of 6.9 years for the paracetamol and 4.4 years for the ibuprofen cohorts. In the primary analyses, the treatment windows were defined as current use ending 3 months after the estimated end of the prescription. In additional analyses by De Vries et al., crude hazard rates were calculated up to 36 months following first exposure.

In the current study, full covariate adjustment had a substantial impact on the risk estimates for all outcomes. Prior publications have presented effect estimates on a variety of adverse outcomes using adjustment methods based on limited selections of baseline variables and likely understated the degree of uncertainty inherent in their conclusions due to residual error in their analytic design. We found that, despite LSPS adjustment and calibration for negative controls, substantial risk of residual error persisted, suggesting the need for caution in the interpretation of nominal statistics. A striking feature of the calibrated CIs is how wide they are. These models are recognizing the variability in the negative control outcome HR estimates and appropriately add uncertainty into the estimates.

4.1 Strengths

As part of the assessment of adverse outcomes in paracetamol use, we also demonstrated the impact of incomplete adjustment for confounding. In prior studies, much of the confounding has been attributed to channeling bias, but the extent of that influence has not been measured. Here we used standardized differences to measure the adequacy of adjustment. The large magnitude in numerous standardized differences in a wider range of available variables after matching on publication covariates indicated important differences in the comparison groups.

This is the first study of adverse outcomes in paracetamol versus ibuprofen to make use of negative controls to evaluate the adequacy of the study design to control for systematic bias. We observed that there was substantial error in the data after controlling for possible confounding through the use of LSPS matching. The magnitude of error contradicts and questions the certainty of conclusions in some prior studies on the risk of paracetamol versus other OTC analgesics.

4.2 Limitations

The data source for this study does not capture all exposures for single-ingredient paracetamol and ibuprofen since small quantity packages are also available OTC. The reasons for a GP to prescribe a medication that is available OTC are discussed in Sect. 2.2 and include patient need for long-term medication as well as issues related to cost. Patients in the study therefore may be either sicker or in a lower income group than those missed, since long-term or lower income users would be more likely to benefit from having the prescription dispensings where there is no cost. Non-differential misclassification of exposure is expected to bias estimates toward 1.0, but the direction of potential exposure misclassification is unknown here. Despite our use of both LSPS matching and negative controls for calibration, we were not able to entirely mitigate the error in the data. Exposure due to OTC use as well as potentially important confounders such as exercise, education, and occupation were not captured in the data.

Furthermore, the censoring on discontinuation of a drug could be informative censoring if the discontinuation is for reasons that are predictive of study outcomes.

Ideally, negative controls would be selected not only for not being impacted by the exposure of interest, but also for being subject to the same mechanism of confounding as might affect the study being conducted. However, we believe the mechanism of confounding is unknowable, and if knowable, such perfect negative controls may not exist. Our selection of negative controls therefore focused on the former criterion and used a large sample of negative controls, assuming that this samples from the types of confounding that could exist in our study.

5 Conclusions

In this comparative cohort study assessing use of propensity score matching in the estimation of risk of MI, stroke, GI bleed, and acute renal failure in patients treated with paracetamol versus ibuprofen, results varied substantially depending on the models used to control for confounding and bias. LSPS matching resulted in attenuated effects and increased precision. After calibration, substantial bias remained, undermining the ability to reasonably discern or rule out an effect of paracetamol exposure on these outcomes. For comparisons of paracetamol versus ibuprofen, when using our methods against the CPRD, it is only possible to distinguish true effects, if those true effects are large (HR > 2), due to residual bias. For the outcomes under study here, where we would expect smaller HRs even if there were a true effect, they cannot be readily differentiated from null effects. In other words, only true effects that rise above the null effects, as estimated by the negative controls, can be detected. Therefore, we cannot conclude that paracetamol increases risk relative to ibuprofen, nor can we conclude it does not, based on these data alone. Future research on adverse health outcomes' association with paracetamol versus ibuprofen should demonstrate adequate adjustment for bias or include a measure of the success of such adjustment.

Compliance with Ethical Standards

Funding No sources of funding were used to conduct this study or prepare this article. Johnson & Johnson will be the sponsor of Open Access, if applicable.

Availability of Data and Material This study is based on data from the CPRD obtained under license from the UK Medicines and Healthcare Products Regulatory Agency.

Code Availability The CPRD has been transformed into the OMOP CDM, version 5. The complete specification for OMOP CDM, version 5, is available at https://github.com/OHDSI/CommonDataModel . The extract, transform, load (ETL) specification for transforming the CPRD into the OMOP CDM is available at https://github.com/OHDSI/ETL-CDMBuilder/tree/master/man/CPRD. The analysis source code for this study was posted on the Observational Health Data Sciences and Informatics (OHDSI) website Repository of OHDSI Collaborative Research Protocols (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ParacetamolvIbuprofen_MI).

Conflict of interest Jesse A. Berlin, Daniel Fife, Patrick B. Ryan, Martijn J. Schuemie, Joel Swerdel, and Rachel B. Weinstein were full-time employees of Johnson & Johnson, or a subsidiary, at the time the study was conducted. Johnson & Johnson manufactures several medications that contain paracetamol or ibuprofen. Jesse A. Berlin, Daniel Fife, Patrick B. Ryan, Martijn J. Schuemie, Joel Swerdel, and Rachel B. Weinstein own stock, stock options, and pension rights from the company.

Ethics approval The pre-specified protocol for this study is available at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ParacetamolvIbuprofen_MI and was approved by the data provider's Independent Scientific Advisory Committee (Protocol 18_100R).

References

1. Roberts E, Delgado Nunes V, Buckner S, Latchem S, Constanti M, Miller P, Doherty M, Zhang W, Birrell F, Porcheret M, Dziedzic K, Bernstein I, Wise E, Conaghan PG. Paracetamol: not as safe as we thought? A systematic literature review of observational studies. Ann Rheum Dis. 2015;0:1–8.
2. Lipworth L, Friis S, Mellemkjaer L, et al. A population-based cohort study of mortality among adults prescribed paracetamol in Denmark. J Clin Epidemiol. 2003;56:796–801.
3. de Vries F, Setakis E, van Staa TP, et al. Concomitant use of ibuprofen and paracetamol and the risk of major clinical safety outcomes. Br J Clin Pharmacol. 2010;70:429–38.
4. Chan AT, Manson JE, Albert CM, et al. Nonsteroidal anti-inflammatory drugs, paracetamol, and the risk of cardiovascular events. Circulation. 2006;113:1578–87.
5. Sandler DP, Smith JC, Weinberg CR, et al. Analgesic use and chronic renal disease. N Engl J Med. 1989;320(19):1238–43.
6. Perneger TV, Whelton PK, Klag MJ. Risk of kidney failure associated with the use of acetaminophen, aspirin, and nonsteroidal antiinflammatory drugs. N Engl J Med. 1994;331(25):1675–9.
7. Kurth T, Glynn RJ, Walker AM, et al. Analgesic use and change in kidney function in apparently healthy men. Am J Kidney Dis. 2003;42:234–44.
8. Evans M, Fored CM, Bellocco R, et al. Acetaminophen, aspirin and progression of advanced chronic kidney disease. Nephrol Dial Transplant. 2009;24:1908–18.
9. Dedier J, Stampfer M, Hankinson S, et al. Nonnarcotic analgesic use and the risk of hypertension in US women. Hypertension. 2002;40:604–8.
10. Curhan GC, Willett WC, Rosner B, et al. Frequency of analgesic use and risk of hypertension in younger women. Arch Intern Med. 2002;162:2204–8.
11. Curhan GC, Knight EL, Rosner B, Hankinson SE, Stampfer MJ. Lifetime nonnarcotic analgesic use and decline in renal function in women. Arch Intern Med. 2004;164(14):1519–24.
12. Fored CM, Ejerblad E, Lindblad P. Acetaminophen, aspirin and chronic renal failure. N Engl J Med. 2001;345:1801–8.
13. Weinstein RB, Ryan P, Berlin JA, et al. Channeling in the use of nonprescription paracetamol and ibuprofen in an electronic medical records database: evidence and implications. Drug Saf. 2017;40(12):1279–92.
14. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2014;33(2):209–18.
15. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci USA. 2018;115(11):2571–7.
16. Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, et al. How confident are we about observational findings in

health care: a benchmark study. Harvard Data Sci Rev. 2020. https://doi.org/10.1162/99608f92.147cc28e.

17. Herret E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827–36.
18. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a Clinical Practice Research Datalink conversion to the OMOP common data model. Drug Saf. 2014;37:945–59.
19. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22(3):553–64. https://doi.org/10.1093/jamia/ocu023.
20. Makadia R, Ryan PB. Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. EGEMS (Wash DC). 2014;2(1):1110. https://doi.org/10.13063/2327-9214.1110.
21. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Trans Model Comput Simul Publ Assoc Comput Mach. 2013;23(1):1–17.
22. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc B. 1996;58(1):267–88.
23. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol. 2015;68:627–6.
24. Walker AM, Patrick AR, Lauer MS, Hornbrook MC, Marin MG, Platt R, Roger VL, Stang P, Schneeweiss S. A tool for assessing the feasibility of comparative effectiveness research. Comp Effect Res. 2013;2013(3):11–20.
25. Dusetzina SB, Brookhart MA, Maciejewski ML. Control outcomes and exposures for improving internal validity of nonrandomized studies. Health Serv Res. 2015;50:1432–51.
26. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology (Cambridge, Mass). 2010;21:383–8.
27. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. J Biomed Inform. 2017;66:72–81.
28. Ryan PB, Buse JB, Schuemie MJ, et al. Comparative effectiveness of canagliflozin, SGLT2 inhibitors and non-SGLT2 inhibitors on the risk of hospitalization for heart failure and amputation in patients with type 2 diabetes mellitus: a real-world meta-analysis of 4 observational databases (OBSERVE-4D). Diabetes Obes Metab. 2018;20(11):2585–97. https://doi.org/10.1111/dom.13424.
29. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet. 2019;394(10211):1816–26. https://doi.org/10.1016/S0140-6736(19)32317-7.
30. Duke JD, Ryan PB, Suchard MA, et al. Risk of angioedema associated with levetiracetam compared with phenytoin: findings of the observational health data sciences and informatics research network. Epilepsia. 2017;58(8):e101–e106106. https://doi.org/10.1111/epi.13828.