Title

Comparative safety and effectiveness of alendronate versus raloxifene in women with osteoporosis

Journal

Scientific Reports, 10(1)

ISSN

2045-2322

Authors

Kim, Yeesuk
Tian, Yuxi
Yang, Jianxiao
et al.

Publication Date

2020

DOI

10.1038/s41598-020-68037-8

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Comparative safety and effectiveness of alendronate versus raloxifene in women with osteoporosis

Yeesuk Kim[1]✉, Yuxi Tian[2], Jianxiao Yang[2], Vojtech Huser[3], Peng Jin[4], Christophe G. Lambert[5], Hojun Park[6], Seng Chan You[6], Rae Woong Park[6], Peter R. Rijnbeek[7], Mui Van Zandt[8], Christian Reich[8], Rohit Vashisht[9], Yonghui Wu[10], Jon Duke[11], George Hripcsak[4,12], David Madigan[13], Nigam H. Shah[9], Patrick B. Ryan[14], Martijn J. Schuemie[14] & Marc A. Suchard[2,15,16]

Alendronate and raloxifene are among the most popular anti-osteoporosis medications. However, there is a lack of head-to-head comparative effectiveness studies comparing the two treatments. We conducted a retrospective large-scale multicenter study encompassing over 300 million patients across nine databases encoded in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The primary outcome was the incidence of osteoporotic hip fracture, while secondary outcomes were vertebral fracture, atypical femoral fracture (AFF), osteonecrosis of the jaw (ONJ), and esophageal cancer. We used propensity score trimming and stratification based on an expansive propensity score model with all pre-treatment patient characteristics. We accounted for unmeasured confounding using negative control outcomes to estimate and adjust for residual systematic bias in each data source. We identified 283,586 alendronate patients and 40,463 raloxifene patients. There were 7.48 hip fracture, 8.18 vertebral fracture, 1.14 AFF, 0.21 esophageal cancer and 0.09 ONJ events per 1,000 person-years in the alendronate cohort and 6.62, 7.36, 0.69, 0.22 and 0.06 events per 1,000 person-years, respectively, in the raloxifene cohort. Alendronate and raloxifene have a similar hip fracture risk (hazard ratio [HR] 1.03, 95% confidence interval [CI] 0.94–1.13), but alendronate users are more likely to have vertebral fractures (HR 1.07, 95% CI 1.01–1.14). Alendronate has higher risk for AFF (HR 1.51, 95% CI 1.23–1.84) but similar risk for esophageal cancer (HR 0.95, 95% CI 0.53–1.70), and ONJ (HR 1.62, 95% CI 0.78–3.34). We demonstrated substantial control of measured confounding by propensity score adjustment, and minimal residual systematic bias through negative control experiments, lending credibility to our effect estimates. Raloxifene is as effective as alendronate and may remain an option in the prevention of osteoporotic fracture.

[1]Department of Orthopaedic Surgery, College of Medicine, Hanyang University, Seoul 04763, Republic of Korea. [2]Department of Computational Medicine, University of California, Los Angeles, CA 90095, USA. [3]Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD 20894, USA. [4]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA. [5]Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA. [6]Department of Biomedical Informatics, Ajou University, Suwon 16499, Republic of Korea. [7]Department of Medical Informatics, Erasmus University Medical Center, 3000 Rotterdam, CA, The Netherlands. [8]Real World Insights, IQVIA, Cambridge, MA 02139, USA. [9]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA. [10]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [11]Center for Health Analytics and Informatics, Georgia Tech Research Institute, Atlanta, GA 30332, USA. [12]Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY 10032, USA. [13]Department of Statistics, Columbia University, New York, NY 10032, USA. [14]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560, USA. [15]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA. [16]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA. ✉email: estone96@gmail.com

Osteoporosis is a chronic, progressive disorder characterized by unbalanced bone resorption, decreased bone mass, and deterioration of the bone microarchitecture, leading to decreased bone strength and increased fracture susceptibility[1,2]. Osteoporosis has substantial disease burden worldwide, and postmenopausal women are especially at risk, with prevalence ranging from approximately 20% in the United States and the European Union to nearly 40% in South Korea and Japan[3–5].

The bisphosphonate alendronate and the selective estrogen receptor modulator (SERM) raloxifene are among the most popular antiresorptive agents for the prevention and treatment of postmenopausal osteoporosis[6,7]. Based on existing randomized studies that compare alendronate and raloxifene separately to placebo[8,9], alendronate seems to have superior fracture prevention benefits. However, few randomized studies evaluate head-to-head comparative effectiveness of osteoporosis drugs that should inform patient treatment decisions[10]. Observational studies can provide evidence missing from the randomized study literature, especially regarding rare but serious adverse events that require large study populations to detect. Two existing observational studies performed propensity score (PS) adjusted comparative effectiveness analysis on insurance claims databases and find no difference in both vertebral and nonvertebral fracture rates between alendronate and raloxifene patients[3,11]. However, they did not address suspected serious adverse events such as atypical femoral fractures (AFF), esophageal cancer, and osteonecrosis of the jaw (ONJ).

In this paper, we leveraged the research network of the Observational Health Data Sciences and Informatics (OHDSI) collaborative[12] to conduct a multicenter retrospective cohort study across nine databases investigating comparative risks of fractures and select adverse events among first-time initiators of alendronate and raloxifene. We implemented a suite of methods to address observational study confounding, including propensity score (PS) adjustment to control for measured confounding and negative control experiments, an emerging observational analytics tool[13], to quantify and adjust for residual study bias.

Methods

Data sources. We conducted a new-user cohort study comparing first-time users of alendronate with new users of raloxifene in nine clinical data sources encoded in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5 from participating research partners across the OHDSI community[12,14,15]. Three data sources were electronic medical records: University of Texas Cerner Health Facts Database (total of 2.4 million [M] patients), Columbia University Medical Center/NewYork-Presbyterian Hospital (4.5M) and Stanford University Hospital (2M). Six data sources are claims records: OptumInsight's Clinformatics Datamart (Eden Prairie, MN) (CEDM, 40.7M), Truven MarketScan Commercial Claims and Encounters (CCAE, 122M), Truven MarketScan Multi-State Medicaid (MDCD, 17.3M), Truven MarketScan Medicare Supplemental Beneficiaries (MDCR, 9.3M), IQVIA PharMetrics Plus (P-Plus, 105M), and the Korean National Health Insurance Service - National Sample Cohort (NHIS NSC, 1.1M). All were mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) schema, providing a homogeneous format for healthcare data and standardization of underlying clinical coding systems that thus enables analysis code to be shared across participating datasets in the network[16,17]. OHDSI network studies are carried out through a federated model, where the access to data and statistical testing executes inside the firewall of the research partners' infrastructure on de-identified patient information, and the research coordinators collect aggregate results absent of patient-level information for meta-analysis, interpretation, and manuscript generation. Each data partner consulted on a shared study design, including all decisions on cohort definitions and statistical methodology, and presentation of results. Afterwards, each data partner executed an identical study package, so there are no differences in study design across databases.

Study design. This study followed a retrospective, observational, comparative cohort design[18]. We included women over 45 years old who were first time users of alendronate or raloxifene from January 2001 to February 2012, and who had a diagnosis of osteoporosis in the year prior to treatment initiation. Patients were required to have continuous observation in the database for at least one year prior to treatment initiation and 90 days after. We excluded patients with a previous diagnosis of hip fracture, high-energy trauma, or other diseases related to pathological fractures (including Paget's disease), as well as patients with prior hip replacements or exposure to any bisphosphonate (including alendronate) or the SERMs raloxifene and bazedoxifene. We used raloxifene as the reference treatment. Full cohort details, including concept codes, are provided in the eMethods in the Supplementary materials.

The primary outcome of interest was osteoporotic hip fracture, while secondary outcomes included vertebral fracture and suspected adverse events: atypical femoral fracture (AFF), osteonecrosis of the jaw (ONJ), and esophageal cancer. We began the outcome risk window at 90 days after treatment initiation, and excluded patients with prior occurrence of that outcome before the risk window. As our primary analysis, we have elected before executing the study to end the outcome time-at-risk window when the patient was no longer observable in the database, analogous to an intent-to-treat design. In addition, to assess the sensitivity of our results to this decision, we considered an alternative analysis in which we ended the time-at-risk window at first cessation of the continuous drug exposure, analogous to an on-treatment design. Continuous drug exposures were constructed from the available longitudinal data by considering sequential prescriptions that had fewer than 30 days gap between prescriptions.

Ethical considerations. The study was conducted in accordance with the rules of the Declaration of Helsinki of 1975, revised in 2013. The use of Optum and Truven Marketscan databases was reviewed by the New England IRB and was determined to be exempt from broad IRB approval, as this research project did not involve human subjects research. This study was approved with waiver of informed consent by the Columbia University

Outcome	Alendronate				Raloxifene			
	Patients	Person-Years	Events	Rate[a]	Patients	Person-Years	Events	Rate[a]
Primary analysis								
Hip fracture	283,586	1,076,597	8051	7.48	40,463	156,080	1033	6.62
Vertebral fracture	279,497	1,058,734	8659	8.18	40,051	154,031	1134	7.36
Atypical femoral fracture	283,894	1,094,049	1244	1.14	40,503	158,722	109	0.69
Esophageal cancer	283,981	1,096,983	234	0.21	40,482	158,858	35	0.22
Osteonecrosis of jaw	284,079	1,097,499	101	0.09	40,511	158,972	9	0.06
Alternative analysis[b]								
Hip fracture	185,021	116,262	622	5.35	27,620	17,282	92	5.32
Vertebral fracture	182,025	114,510	719	6.28	27,308	17,066	112	6.56
Atypical femoral fracture	185,258	116,735	85	0.73	27,642	17,345	6	0.35
Esophageal cancer	185,312	116,801	13	0.11	27,625	17,348	< 6	0.23
Osteonecrosis of jaw	185,367	116,838	< 6	0.03	27,649	17,365	0	0

Table 1. Size of study cohorts for each outcome of interest in primary and alternative analyses. [a]Rate: incidence per 1,000 person-years. [b]Three data sources excluded from alternative analysis (P-Plus, NHIS NSC, Cerner UT).

Institutional Review Board under protocol IRB-AAAO7805, most recently renewed 6/11/2019. The research at Stanford was reviewed by their Administrative Panels for the Protection of Human Subjects under protocols 24883 to obtain de-identified data and 53248 to participate in OHDSI network studies. The IRB of Ajou University, Republic of Korea approved the research(AJIRB-MED-EXP-17-24).

Statistical analysis. We conducted our cohort study using the open-source OHDSI CohortMethod R package[19], with large-scale analytics achieved through the Cyclops R package[20]. We used propensity scores (PSs)—estimates of treatment exposure probability conditional on pre-treatment baseline features in the 1 year prior to treatment initiation—to control for potential confounding and improve balance between the target (alendronate) and reference (raloxifene) cohorts[21]. We used an expansive PS model that includes all available patient demographic, drug, condition, and procedure covariates instead of a prespecified set of investigator-selected confounders[22]. We performed PS trimming and stratification and then estimated comparative alendronate-vs-raloxifene hazard ratios (HR) using a Cox proportional hazards model. Detailed covariate and methods information are provided in the eMethods in the Supplementary material. We presented PS and covariate balance metrics to assess successful confounding control, and provided hazard ratio estimates and Kaplan–Meier survival plots for the outcomes of interest.

Residual study bias from unmeasured and systematic sources can exist in observational studies after controlling for measured confounding. To estimate such residual bias, we conducted negative control outcome experiments with 147 negative control outcomes[23], identified through a data-rich algorithm[24]. Negative control outcomes, separate of our study outcomes, are events believed to be unaffected by the studied treatments, thus having a presumed true HR of 1 (See the eMethods in the Supplementary materials for the list of included negative controls). We fitted the negative control estimates to an empirical null distribution that characterizes the study residual bias and is an important artifact from which to assess the study design[25].

Results

Population characteristics. Across all data sources, we identifed 283,586 alendronate patients and 40,463 raloxifene patients for the primary hip fracture analysis, totaling 1,076,597 and 156,080 patient-years of observation, respectively; corresponding cohort sizes for all study outcomes were similar (Table 1). Approximately 98% of patients came from claims databases, and two Electronic Health Records(EHRs)—Columbia and Stanford—had very low numbers of raloxifene users and contributed only modest information (eTable 1 in Supplementary material). The data sources showed a diversity of study entry year and age at study entry distributions (eFigure 1 in Supplementary material). The on-treatment alternative analysis yielded similar cohort sizes for included data sources (eTable 2 in Supplementary material). However, we excluded three data sources (P-Plus, Cerner UT, NHIS NSC) because of continuous drug era encoding difficulties.

Primary outcome assessment. In the primary analysis, there were 8,051 hip fractures out of 283,586 patients and the unadjusted rate was 7.48 hip fracture per 1,000 person-years in alendronate. In the raloxifene group, 1,033 out of 40,463 patients had hip fractures with an unadjusted rate of 6.62 hip fractures per 1,000 person-years. The rates in the on-treatment alternative analysis were 5.35 for alendronate and 5.32 for raloxifene (Table 1). Neither the primary analysis across all data sources (summary HR 1.03, 95% CI 0.94–1.13) (Fig. 1a) nor the on-treatment alternative (summary HR 0.88, 95% CI 0.71–1.11) (Fig. 1b) demonstrated a statistically significant difference between treatments. Kaplan-Meier plots across data sources showed small differences between alendronate and raloxifene survival (eFigure 2 in Supplementary material).

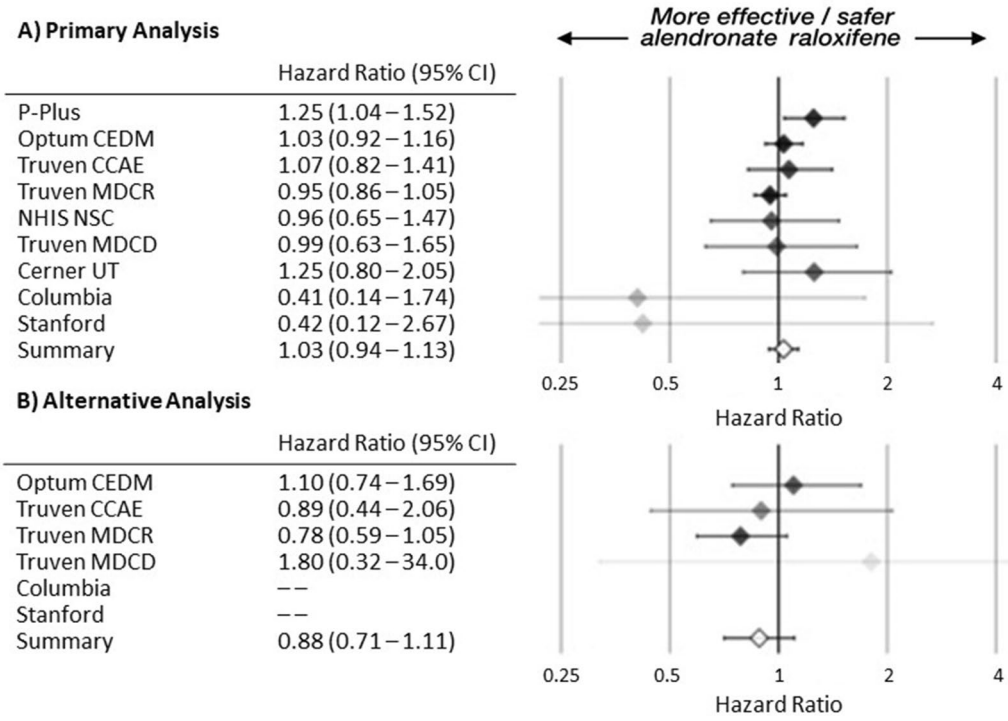

Fig. 1. (**A**) Primary and (**B**) alternative analysis hazard ratios (HRs) for hip fracture. More precise estimates have greater opacity. Missing HR from data source with 0 raloxifene events.

Secondary outcome assessment. In the primary analysis, there were 8.18 vertebral fracture, 1.14 AFF, 0.21 esophageal cancer, and 0.09 ONJ outcomes per 1,000 person-years in the alendronate cohort, compared to 7.36, 0.69, 0.22 and 0.06, respectively, in the raloxifene cohort (Table 1). Alendronate users are more likely to have vertebral fractures (summary HR 1.07, 95% CI 1.01–1.14) (Fig. 2a), and have higher risk for AFF (summary HR 1.51, 95% CI 1.23–1.84) (Fig. 2b). There was no significant difference in esophageal cancer risk (summary HR 0.95, 95% CI 0.53–1.70) (Fig. 2c), or ONJ risk (summary HR 1.62, 95% CI 0.78–3.34) (Fig. 2d).

In the on-treatment alternative analysis, the respective rates for the four secondary outcomes were 6.28, 0.73, 0.11, 0.03 among alendronate users and 6.56, 0.35, 0.23, 0.00 among raloxifene users (Table 1). Some data sources had 0 events among one or both treatment groups, and consequently had nonexistent HR estimates. We found no significant vertebral fracture risk (summary HR 0.87, 95% CI 0.71–1.07) and losed power in the other three hypotheses, with extremely wide confidence intervals for AFF and esophageal cancer and 0 raloxifene cohort outcomes for ONJ (eFigure 3 in Supplementary material).

Cohort balance. Across all data sources, preference score distributions, re-scalings of PS estimates to adjust for differential treatment prevalences, were generally similar and have large overlap between treatment groups, suggesting a meaningful comparative effectiveness study (eFigure 4 in Supplementary material). A large majority of patients had intermediate preference scores, and all data sources except Cerner UT and NHIS NSC displayed at most 10% loss to preference trimming to 0.25–0.75 (Table 2).

We assessed the covariate balance achieved through PS adjustment by comparing all covariates' standardized mean difference between treatment groups before and after PS trimming and stratification, as shown graphically for all data sources (eFigure 5 in Supplementary material), with summary statistics for all data sources shown in Table 3. In all but one data source (Stanford) that had poor PS differentiation, there were large decreases from PS adjustment in both the standardized mean difference and the proportion of covariates with standardized mean difference greater than 0.05. For example, in the P-Plus database, the raloxifene-related covariate "gynecologic examination" is the most unbalanced pre-adjustment covariate, with a standardized mean difference over 0.2. After PS trimming and stratification, this and all other covariates have standardized mean differences smaller than 0.05, indicating successful balancing (Fig. 3).

Considering the top unbalanced covariates in the adjusted cohorts, alendronate users had more bone disorders (eFigures 6–14 in Supplementary material). Meanwhile, raloxifene users had more gastrointestinal counterindications to alendronate and gynecologic examinations and procedures due to its alternate use for breast cancer treatment. These clinical covariates were expected potential confounders and across all data sources became successfully balanced through PS adjustment, reducing the bias in our effect estimates.

Negative control outcomes. In the absence of bias, 95% of the negative control estimates' 95% confidence intervals were expected to include the presumed null HR of 1. Across data sources, the proportion of

A) Vert. Fracture

	Hazard Ratio (95% CI)
P-Plus	1.02 (0.88 – 1.18)
Optum CEDM	1.05 (0.94 – 1.17)
Truven CCAE	1.19 (0.97 – 1.49)
Truven MDCR	1.10 (0.99 – 1.22)
NHIS NSC	1.06 (0.37 – 4.47)
Truven MDCD	0.98 (0.62 – 1.65)
Cerner UT	1.09 (0.72 – 1.69)
Columbia	– –
Stanford	1.32 (0.28 – 23.7)
Summary	1.07 (1.01 – 1.14)

B) AFF

	Hazard Ratio (95% CI)
P-Plus	1.72 (1.03 – 3.12)
Optum CEDM	1.40 (1.01 – 1.99)
Truven CCAE	1.62 (0.83 – 3.66)
Truven MDCR	1.56 (1.13 – 2.21)
NHIS NSC	1.45 (0.65 – 4.16)
Truven MDCD	0.93 (0.32 – 3.92)
Cerner UT	1.99 (0.66 – 8.64)
Columbia	– –
Stanford	– –
Summary	1.51 (1.23 – 1.84)

C) Esophageal Cancer

	Hazard Ratio (95% CI)
P-Plus	0.46 (0.26 – 0.89)
Optum CEDM	1.12 (0.58 – 2.44)
Truven CCAE	0.83 (0.39 – 2.06)
Truven MDCR	2.42 (0.99 – 8.00)
NHIS NSC	1.38 (0.27 – 25.4)
Truven MDCD	– –
Cerner UT	– –
Columbia	– –
Stanford	– –
Summary	0.95 (0.53 – 1.70)

D) ONJ

	Hazard Ratio (95% CI)
P-Plus	2.63 (0.79 – 16.3)
Optum CEDM	1.84 (0.65 – 7.68)
Truven CCAE	0.99 (0.33 – 4.24)
Truven MDCR	1.68 (0.31 – 31.1)
NHIS NSC	– –
Truven MDCD	– –
Cerner UT	– –
Columbia	– –
Stanford	– –
Summary	1.62 (0.78 – 3.34)

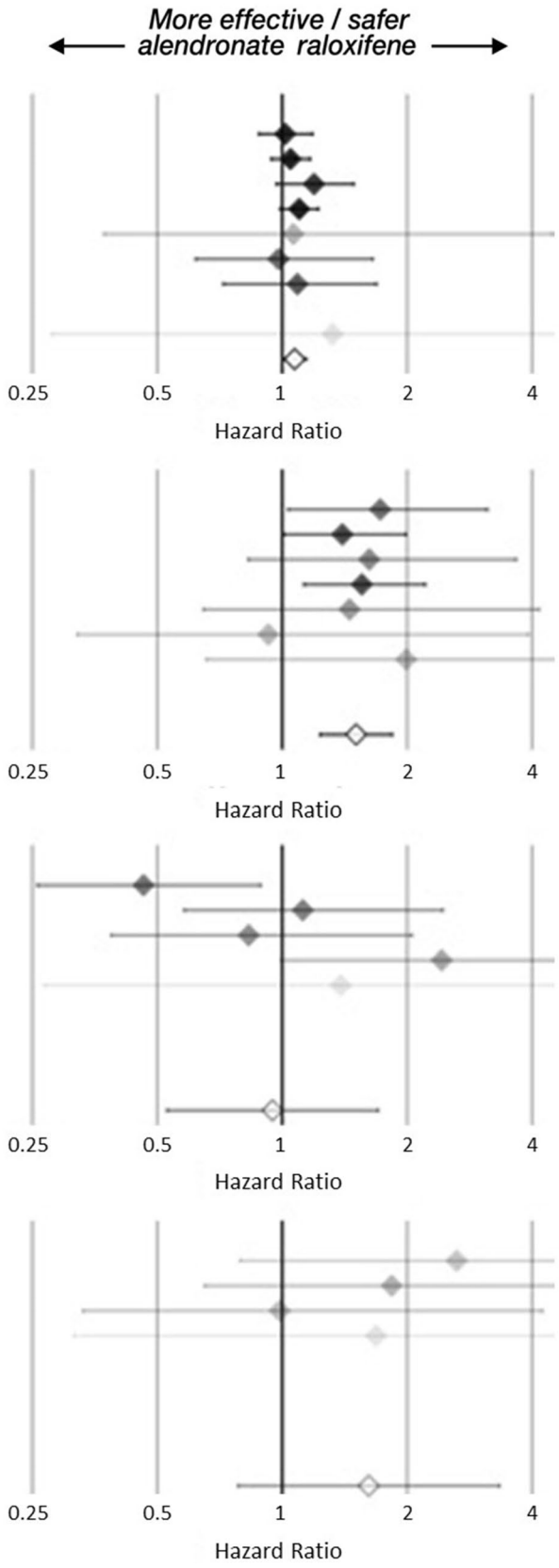

Fig. 2. Primary analysis hazard ratios for (**A**) vertebral fracture (Vert. Fracture), (**B**) atypical femoral fracture (AFF), (**C**) esophageal cancer (Eso. Cancer), and (**D**) osteonecrosis of jaw (ONJ). More precise estimates have greater opacity. Missing HR from data sources with 0 raloxifene events.

Data source	Alendronate (%)	Raloxifene (%)	Total (%)
P-Plus	10	11	10
Optum CEDM	7.20	5.90	7
Truven CCAE	3.40	3.80	3.40
Truven MDCR	5.80	6.50	5.90
NHIS NSC	12	17	13
Truven MDCD	7.90	14.00	8.40
Cerner UT	21	19	21
Columbia	0	0	0
Stanford	0	0	0

Table 2. Percentage of cohort eliminated by trimming to 0.25–0.75 preference score.

		Before PS		After PS	
Data source	Covariates	Mean	> 0.05(%)	Mean	> 0.05(%)
P-Plus	6611	0.23	6.1	0.04	0
Optum CEDM	6890	0.2	8.2	0.05	0.015
Truven CCAE	5605	0.16	4.3	0.05	0
Truven MDCR	4726	0.2	8.8	0.06	0.11
NHIS NSC	3138	0.36	26	0.13	21
Truven MDCD	1873	0.32	53	0.21	49
Cerner UT	721	0.46	72	0.13	20
Columbia	379	0.73	84	0.44	66
Stanford	288	0.45	81	0.44	81

Table 3. Number of covariates by data source, along with mean standardized differences and percentage with standardized difference greater than 0.05 before and after propensity score(PS) adjustment.

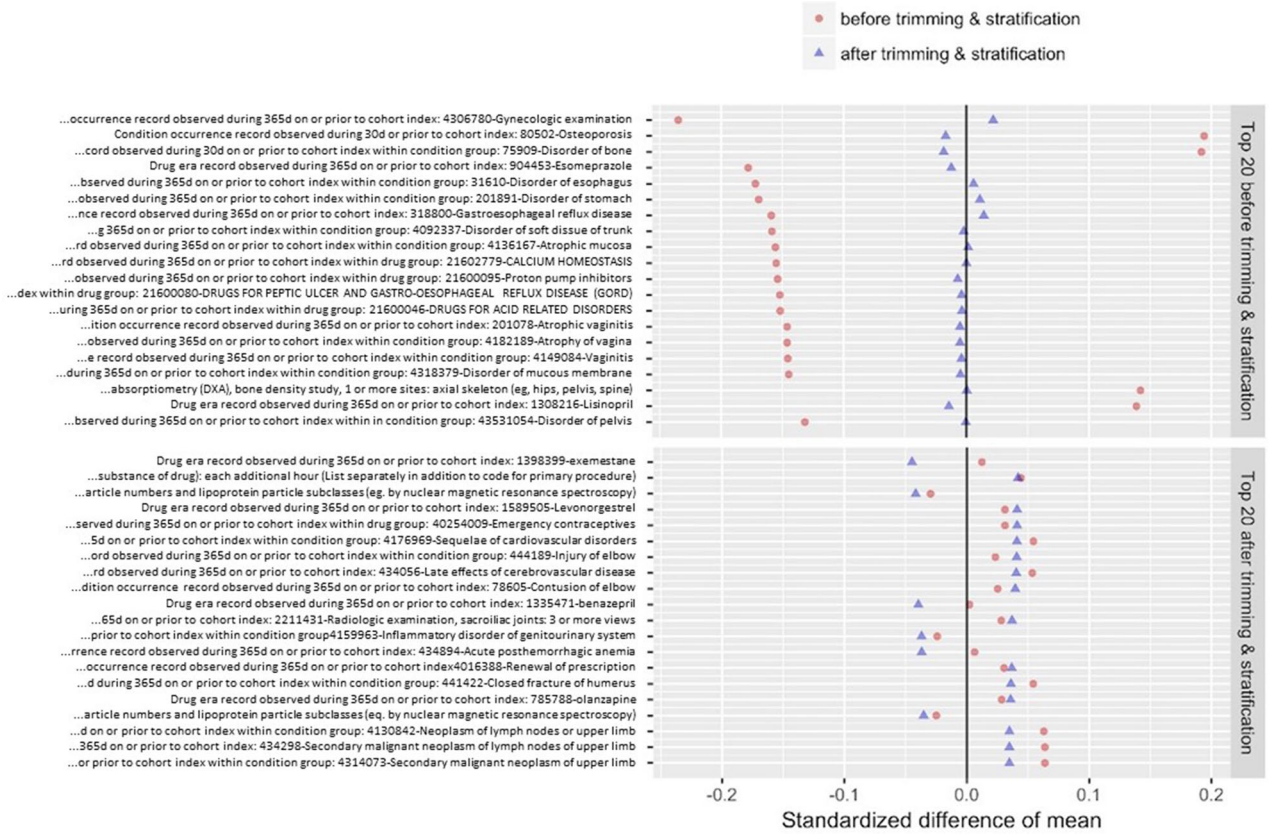

Fig. 3. P-Plus: most unbalanced covariates before (top) and after (bottom) PS trimming and stratification.

estimates that included 1 was high, ranging from 91 to 96% in primary analysis (eTables 3, 4 in Supplementary material). Furthermore, Gaussian empirical null distributions that estimated the residual bias were centered close to 1 for all data sources except the Columbia and Stanford EHRs that had few raloxifene patients (eTables 5, 6 in Supplementary material). As a result, the theoretical (uncalibrated) and negative control calibrated p-value distributions were very similar to each other (eFigures 15–18 in Supplementary material). These results indicated minimal residual bias across data sources for both primary and alternative analyses, giving further credence to the relative unbiasedness of our treatment effect estimates.

Discussion

Prevailing clinical wisdom favors alendronate as the first-line treatment option for osteoporosis patients against fracture[26–30]. However, head-to-head randomized studies of alendronate vs raloxifene have only shown increased bone mineral density with alendronate[31,32], which do not necessarily relate to clinically observed fracture risk[6,9]. Our results found little difference in hip fracture risk between new users of alendronate and raloxifene, and also found a small but statistically significant higher vertebral fracture risk with alendronate. Foster et al. reported non-significantly higher alendronate vertebral fracture risk compared to raloxifene using Truven CCAE and Truven MDCR data[3]. Our data sources were individually similarly non-significant, but together they provided the requisite population size to reveal a statistically significant effect favoring raloxifene.

Growing concern over long-term bisphosphonate use has contributed to steep declines in their prescription[33]. Previous studies report conflicting non-significant[8,34] and positively significant[35,36] estimates for atypical femoral fracture risk as a result of bisphosphonate-related suppression of bone remodeling[37]. We found that compared to raloxifene, alendronate did lead to increased AFF risk. Importantly, this well-known and statistically significant risk difference demonstrated that our data sources and study design furnished sufficient statistical power to detect a true difference in the hip fracture HR if one were to exist, given that the rates of AFF were almost an order-of-magnitude less than of hip fracture in our data.

Further, upper gastrointestinal mucosa stimulation is a common bisphosphonate adverse event[38–41], and concern of bisphosphonate related esophageal cancer has been discussed. Specifically, the US Food and Drug Administration (FDA) received reports of 23 esophageal cancer patients who have taken the alendronate as a suspect drug[42]. After then, the report from the UK primary care cohort concluded that the risk of esophageal cancer increased in patients with oral bisphosphonate compared with non-prescriptions[43]. However, in the following reports, association with the related esophageal cancer is less established[44–46]. We found similar esophageal cancer incidence between alendronate and raloxifene users, and no difference in hazard ratio, and similarly found no difference for osteonecrosis of the jaw, although our study was likely underpowered for this very rare adverse event.

It is known that alendronate related adverse effects such as AFF and ONJ may be affected by the duration of drug use. However, studies to determine duration of use and dosage are lacking. With our study design, we could not establish a dose and duration criterion for analysis. However, our alternative on-treatment analysis was able to measure continuous exposure to treatment, and provide average patient-years of exposure between alendronate and raloxifene cohorts. As calculated from Table 1, for the AFF analysis, alendronate users averaged 0.630 patient-years of continuous exposure, while raloxifene users averaged 0.627 patient-years. For the ONJ analysis, alendronate users averaged 0.630 patient-years of continuous exposure, while raloxifene users averaged 0.628 patient-years of continuous exposure. These are extremely small differences in average exposure duration, giving us confidence in our comparative results for AFF and ONJ.

Many sources of bias unique to retrospective, non-randomized data require attention in order to confidently interpret observational study results. Results may vary from database to database because of differences in study population, and the generalizability of a single study is low[47]. However, due to differences in study implementation, results from different studies often cannot be directly compared. Our study, conducted through the OHDSI community, benefits from a large aggregate study population (over 300,000 patients) and standardized data vocabulary, research protocol, and study implementation. To address confounding due to nonrandom treatment assignment present in all observational studies, we performed PS adjustment using an expansive PS model[22] that contrasts with the predominant yet inconsistent and potentially biased approach of hand-selecting covariates[48]. We demonstrated substantial improvements in covariate balance from our PS stratification, including balance of covariates related to bone disease severity, alendronate counterindications, and raloxifene's alternative gynecologic indication.

Meta-analyses often report entirely non-overlapping confidence intervals from different studies investigating the same clinical question. Reported confidence intervals only capture the element of random error, which becomes smaller with larger sample size, but not nonrandom error including study population differences, heterogeneous measurement error, implementation discrepancies, and systematic differences between data sources. Without addressing nonrandom error, divergent study results cannot be reliably combined to leverage the larger aggregate sample sizes across studies. In addition to demonstrating confounding control and using standard research protocols, our study addressed systematic error in each data source through negative control analyses. We use negative controls to quantify systematic bias for this alendronate vs raloxifene comparative effectiveness study, and use the empirical null distribution of negative control estimates to adjust the individual study p-values for our actual outcomes of interest. In this study, we found minimal systematic bias across data sources, providing credibility to our meta-analysis summary hazard ratio estimates.

Our study carries several limitations. Bias from measured and unmeasured sources cannot be ruled out of any observational study, this one included. Data derived from electronic medical records and insurance claims are naturally noisy with missing and misclassified values, and unknown patient histories prior to database entry; our negative control experiments are just one approach to address systematic study bias. Additionally, several

of our insurance claims data sources provide much larger study populations that proportionately dominate the smaller data sources in the meta-analysis. As electronic medical records differ in fundamental ways from claims databases, either separate analyses or more complex meta-analysis weighting schemes may accentuate their unique differences. Having said that, several of our participating electronic medical record data sources have very little treatment or outcome data, and may not be as suitable for comparative effectiveness studies. Usually, alendronate has a strong anti-osteoporotic effect and is the preferred treatment for patients with severe osteoporosis. Unfortunately, the severity of osteoporosis as measured through, for example, low bone mineral density scores, was difficult to assess in claim data alone. However, when reviewing unbalanced covariates before and after PS trimming and stratification in our three electric medical records databases (eFigures 12–14 in Supplementary material), we did not find unbalanced covariates representing severe osteoporotic status. We believe it is unlikely that we successfully balance all measured confounders without having balanced osteoporosis severity, despite the lack of mineral density scores such as the T score in our data. Finally, thromboembolism is a concerning adverse effect of raloxifene. The safety of raloxifene in the treatment of osteoporosis was assessed in a large (7,705 patients) multinational, placebo-controlled trial, and during an average of study-drug exposure of 2.6 years, thromboembolism occurred in about 1 out of 100 patients[9]. Recent 2017 guidelines from the American College of Physicians assert that raloxifene should not be used in osteoporotic women due to cerebrovascular and thromboembolic event concerns[49]. With such concern, raloxifene remains a poor treatment choice for patients with high probability of thromboembolic crises. These guidelines are likely to lead to significant channeling bias, in which patients with a high probability of crisis are very unlikely to receive raloxifene treatment, spuriously inflating the unadjusted rate of thromboembolic events among alendronate users. Such strong bias remains difficult to adjust for using PS modeling alone, as there could be little overlap in probability of treatment across high-risk patients. As a consequence, we are unable to report on the relative hazards of thromboembolic events directly.

In our retrospective, head-to-head comparative effectiveness study across nine data sources with common data model, we found that raloxifene was as effective as alendronate, and raloxifene may remain an option in prevention of osteoporotic fracture.

Data availability

The data that support the findings of this study are available from the OHDSI study, but restrictions apply to their availability. These data were used under license for the current study and so are not publicly available. The outcome data and codes are, however, available from the authors upon reasonable request and with permission of the OHDSI study.

Received: 26 January 2020; Accepted: 16 June 2020
Published online: 06 July 2020

References

1. Bone, H. G. *et al.* Ten years' experience with alendronate for osteoporosis in postmenopausal women. *N. Engl. J. Med.* **350**, 1189–1199. https://doi.org/10.1056/NEJMoa030897 (2004).
2. Nguyen, B., Hoshino, H., Togawa, D. & Matsuyama, Y. Cortical thickness index of the proximal femur: A radiographic parameter for preliminary assessment of bone mineral density and osteoporosis status in the age 50 years and over population. *Clin. Orthop. Surg.* **10**, 149–56. https://doi.org/10.4055/cios.2018.10.2.149 (2018).
3. Foster, S. A. *et al.* Fractures in women treated with raloxifene or alendronate: A retrospective database analysis. *BMC Womens Health* **13**, 15. https://doi.org/10.1186/1472-6874-13-15 (2013).
4. Hernlund, E. *et al.* Osteoporosis in the European Union: Medical management, epidemiology and economic burden. A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA). *Arch Osteop.* **8**, 136. https://doi.org/10.1007/s11657-013-0136-1 (2013).
5. Park, E. J. *et al.* Prevalence of osteoporosis in the Korean population based on Korea National Health and Nutrition Examination Survey (KNHANES), 2008–2011. *Yonsei Med. J.* **55**, 1049–57. https://doi.org/10.3349/ymj.2014.55.4.1049 (2014).
6. Lin, T. *et al.* Alendronate versus raloxifene for postmenopausal women: A meta-analysis of seven head-to-head randomized controlled trials. *Int. J. Endocrinol.* **2014**, 796510. https://doi.org/10.1155/2014/796510 (2014).
7. Miller, P. D. & Derman, R. J. What is the best balance of benefits and risks among anti-resorptive therapies for postmenopausal osteoporosis?. *Osteop. Int.* **21**, 1793–802. https://doi.org/10.1007/s00198-010-1208-3 (2010).
8. Black, D. M. *et al.* Bisphosphonates and fractures of the subtrochanteric or diaphyseal femur. *N. Engl. J. Med* **362**, 1761–71. https://doi.org/10.1056/NEJMoa1001086 (2010).
9. Ettinger, B. *et al.* Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: Results from a 3-year randomized clinical trial. multiple outcomes of raloxifene evaluation (more) investigators. *JAMA* **282**, 637–45 (1999).
10. Cadarette, S. M. *et al.* Relative effectiveness of osteoporosis drugs for preventing nonvertebral fracture. *Ann. Intern. Med.* **148**, 637–46 (2008).
11. Tanaka, S., Yamamoto, T., Oda, E., Nakamura, M. & Fujiwara, S. Real-world evidence of raloxifene versus alendronate in preventing non-vertebral fractures in Japanese women with osteoporosis: Retrospective analysis of a hospital claims database. *J. Bone Miner. Metab.* **36**, 87–94. https://doi.org/10.1007/s00774-016-0809-0 (2018).
12. Hripcsak, G. *et al.* Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–8 (2015).
13. Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A. & Madigan, D. Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Stat. Med.* **33**, 209–18. https://doi.org/10.1002/sim.5925 (2014).
14. Ryan, P. B., Schuemie, M. J., Gruber, S., Zorych, I. & Madigan, D. Empirical performance of a new user cohort method: Lessons for developing a risk identification and analysis system. *Drug Saf.* **36**(Suppl 1), S59-72. https://doi.org/10.1007/s40264-013-0099-6 (2013).
15. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G. & Stang, P. E. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.* **19**, 54–60. https://doi.org/10.1136/amiajnl-2011-000376 (2012).
16. FitzHenry, F. *et al.* Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. *Appl. Clin. Inform.* **6**, 536–547 (2015).
17. Suchard, M. A. *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis. *Lancet* **394**, 1816–1826 (2019).

18. Ryan, P. Statistical challenges in systematic evidence generation through analysis of observational healthcare data networks. *Stat. Methods Med. Res.* **22**, 3–6. https://doi.org/10.1177/0962280211403601 (2013).
19. Schuemie, M. J., Suchard, M. A. & Ryan, P. B. Cohortmethod: New-user cohort method with large scale propensity and outcome models. https://github.com/OHDSI/CohortMethod (2015). Accessed 21 June 2020.
20. Suchard, M. A., Simpson, S. E., Zorych, I., Ryan, P. & Madigan, D. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans. Model Comput. Simul.*https://doi.org/10.1145/2414416.2414791 *(2013).*
21. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score inobservational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
22. Tian, Y., Schuemie, M. J. & Suchard, M. A. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int. J. Epidemiol.* **47**, 2005–2014. https://doi.org/10.1093/ije/dyy120 (2018).
23. Arnold, B. F. & Ercumen, A. Negative control outcomes: A tool to detect bias in randomized trials. *JAMA* **316**, 2597–2598. https://doi.org/10.1001/jama.2016.17700 (2016).
24. Voss, E. A. *et al.* Accuracy of an automated knowledge base for identifying drug adverse reactions. *J. Biomed. Inform.* **66**, 72–81. https://doi.org/10.1016/j.jbi.2016.12.005 (2017).
25. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences* **201708282**, (2018).
26. Ensrud, K. E. *et al.* Effects of raloxifene on fracture risk in postmenopausal women: The raloxifene use for the heart trial. *J. Bone Miner. Res.* **23**, 112–20. https://doi.org/10.1359/jbmr.070904 (2008).
27. Khosla, S. Increasing options for the treatment of osteoporosis. *N. Engl. J. Med.* **361**, 818–20. https://doi.org/10.1056/NEJMe0905480 (2009).
28. MacLean, C. *et al.* Systematic review: Comparative effectiveness of treatments to prevent fractures in men and women with low bone density or osteoporosis. *Ann. Intern. Med.* **148**, 197–213 (2008).
29. Murad, M. H. *et al.* Clinical review. Comparative effectiveness of drug treatments to prevent fragility fractures: A systematic review and network meta-analysis. *J. Clin. Endocrinol. Metab.* **97**, 1871–80. https://doi.org/10.1210/jc.2011-3060 (2012).
30. Wells, G. A. *et al.* Alendronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. *Cochrane Datab. Syst. Rev.*https://doi.org/10.1002/14651858.CD001155.pub2 *(2008).*
31. Luckey, M. *et al.* Once-weekly alendronate 70 mg and raloxifene 60 mg daily in the treatment of postmenopausal osteoporosis. *Menopause* **11**, 405–15 (2004).
32. Sambrook, P. N. *et al.* Alendronate produces greater effects than raloxifene on bone density and bone turnover in postmenopausal women with low bone density: Results of effect (efficacy of fosamax versus evista comparison trial) international. *J. Intern. Med.* **255**, 503–11. https://doi.org/10.1111/j.1365-2796.2004.01317.x (2004).
33. Wysowski, D. K. & Greene, P. Trends in osteoporosis treatment with oral and intravenous bisphosphonates in the United States, 2002–2012. *Bone* **57**, 423–8. https://doi.org/10.1016/j.bone.2013.09.008 (2013).
34. Kim, S. Y., Schneeweiss, S., Katz, J. N., Levin, R. & Solomon, D. H. Oral bisphosphonates and risk of subtrochanteric or diaphyseal femur fractures in a population-based cohort. *J. Bone. Miner. Res.* **26**, 993–1001. https://doi.org/10.1002/jbmr.288 (2011).
35. Gedmintas, L., Solomon, D. H. & Kim, S. C. Bisphosphonates and risk of subtrochanteric, femoral shaft, and atypical femur fracture: A systematic review and meta-analysis. *J. Bone. Miner. Res.* **28**, 1729–37. https://doi.org/10.1002/jbmr.1893 (2013).
36. Schilcher, J., Michaelsson, K. & Aspenberg, P. Bisphosphonate use and atypical fractures of the femoral shaft. *N. Engl. J. Med.* **364**, 1728–37. https://doi.org/10.1056/NEJMoa1010650 (2011).
37. Shane, E. *et al.* Atypical subtrochanteric and diaphyseal femoral fractures: Report of a task force of the American Society for Bone and Mineral Research. *J. Bone Miner. Res.* **25**, 2267–94. https://doi.org/10.1002/jbmr.253 (2010).
38. Abdelmalek, M. F. & Douglas, D. D. Alendronate-induced ulcerative esophagitis. *Am. J. Gastroenterol.* **91**, 1282–3 (1996).
39. Castell, D. O. "Pill esophagitis"—The case of alendronate. *N. Engl. J. Med.* **335**, 1058–1059. https://doi.org/10.1056/NEJM199610033351412 (1996).
40. de Groen, P. C. *et al.* Esophagitis associated with the use of alendronate. *N. Engl. J. Med.* **335**, 1016–21. https://doi.org/10.1056/NEJM199610033351403 (1996).
41. Liberman, U. I. & Hirsch, L. J. Esophagitis and alendronate. *N. Engl. J. Med.* **335**, 1069–70. https://doi.org/10.1056/NEJM199610033351416 (1996).
42. Wysowski, D. K. Reports of esophageal cancer with oral bisphosphonate use. *N. Engl. J. Med.* **360**, 89–90. https://doi.org/10.1056/NEJMc0808738 (2009).
43. Green, J. *et al.* Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: Case–control analysis within a UK primary care cohort. *BMJ* **341**, c4444. https://doi.org/10.1136/bmj.c4444 (2010).
44. Chen, L. X. *et al.* The carcinogenicity of alendronate in patients with osteoporosis: Evidence from cohort studies. *PLoS One* **10**, e0123080. https://doi.org/10.1371/journal.pone.0123080 (2015).
45. Seo, G. H. & Choi, H. J. Oral bisphosphonate and risk of esophageal cancer: A nationwide claim study. *J Bone Metab* **22**, 77–81. https://doi.org/10.11005/jbm.2015.22.2.77 (2015).
46. Sun, K., Liu, J. M., Sun, H. X., Lu, N. & Ning, G. Bisphosphonate treatment and risk of esophageal cancer: A meta-analysis of observational studies. *Osteoporos Int* **24**, 279–86. https://doi.org/10.1007/s00198-012-2158-8 (2013).
47. Madigan, D. *et al.* Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol* **178**, 645–51. https://doi.org/10.1093/aje/kwt010 (2013).
48. King, G. & Nielsen, R. Why propensity scores should not be used for matching. *Political Analysis* **27**, 435–454 (2019).
49. Qaseem, A., Forciea, M. A., McLean, R. M. & Denberg, T. D. Treatment of low bone density or osteoporosis to prevent fractures in men and women: A clinical practice guideline update from the american college of physicians. *Ann Intern Med* **166**, 818–839. https://doi.org/10.7326/M15-1361 (2017).

Acknowledgements

The analysis is based in part on work from the Observational Health Sciences and Informatics collaborative. OHDSI (http://ohdsi.org) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics. PBR and MJS are employees of Janssen Research & Development. MVZ and CR are employees of IQVIA. PRR has received a research Grant from Janssen Research & Development. MAS has received a contract Grant from Janssen Research & Development. This work was partially supported through National Science Foundation Grant IIS 1251151 (MAS), National Library of Medicine Grant 1F31LM012636-01 (YT), Paul and Daisy Soros Fellowships for New Americans (YT), National Institute of General Medical Sciences R01 GM101430 (NHS, RV). RWP acknowledges support from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (Grant number HI16C0992). This study used NHIS-NSC data (HI16C0992) made by National Health Insurance Service (NHIS). The authors declare no conflict of interest with NHIS. YK acknowledges this research was supported by a grant from the Bio Industrial Strategic Technology

Development Program (20001234) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute(KHIDI), funded by the ministry of Health & Welfare, Republic of Korea (Grant number: HI19C0218).

Author contributions

Y.K., P.B.R., M.J.S., and M.A.S. contributed to the conception and design of the study. Y.K., Y.T., J.Y., P.B.R., M.J.S., and M.A.S. conducted this study. Data collection was conducted by V.H., P.J., C.G.L., H.P., R.W.P., P.R.R., M.V.Z., R.V., Y.W., S.C.Y., J.D., G.H., D.M., C.R., and N.H.S., and data analysis was done by Y.T., M.J.S., and M.A.S. Data interpretation was done by Y.K., Y.T., P.B.R., M.J.S., and M.A.S. Y.K. and Y.T. drafted the manuscript and revising manuscript content was done by Y.K., Y.T. and M.A.S. Approving final version of manuscript was performed by all authors. M.A.S. takes responsibility for the integrity of the data analysis.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68037-8.

Correspondence and requests for materials should be addressed to Y.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.