Drug safety analysis: No sex differences in denosumab

Mickado Codrington, master student Biology of Disease, Graduate School of Life Sciences, Utrecht University, the Netherlands & Dutch Medicines Evaluation Board (CBG-MEB)

Patrick Vrijlandt, MD, Medicines Evaluation Board & University Medical Centre, Groningen, the Netherlands

Layman's summary

Introduction: There are concerns in society about the inclusion of women in medical research. An overrepresentation of men in the developmental process of medicines could likely result in unsafe medical treatment for women. In this research we are looking into the side effects observed during treatment with the drug denosumab. We aim to assess if there is a difference in the incidence of side effects between men and women. Denosumab is a medicine used to treat osteoporosis, a disease that causes bone fragility and risk of fracture, and will be assessed as an example case.

Methods: we compared the incidence of side effects between 4 groups: men and women, each treated with either denosumab or with an inactive treatment. A standard way to assess differences between men and women does not exist, because differences are known to occur with denosumab and inactive treatment, which prevents easy interpretation of side effect incidences. We collected information about side effects from multiple researches that studied the workings of denosumab in osteoporosis patients. Information was retrieved from public sources and the Dutch regulatory authority's database (this is medical evaluation board that improves, reviews and permits the development and use of drugs in the Netherlands). Several visual summaries were made for exploration of the side effect incidences in each group, followed by three different ways statistical analysis.

Results: Publicly available data lacked detail for these analyses. The regulatory authority database provided workable data from 8558 females and 1927 males for analysis. Analyses showed that several side effects had a difference in incidence when comparing men and women (for example: side effects related to the kidneys, muscles and skeleton). A considerably smaller number of side effects displayed a difference in incidence when comparing denosumab and inactive treatment (side effects related to the ear and medical procedure complications). None of the visual summaries and statistical analysis methods showed that the incidence of side effects was affected by sex and type of treatment at the same time. We constructed a new visual summary: the 2-factor efficacy and safety scatter ($2FES_2$), enabling identification side effects of which the incidence is affected by sex, treatment and side effects that display sex differences that are attributable to denosumab.

Conclusion: Statistical analysis does not reveal important sex differences in the incidence of side effects that are attributable to the presence of denosumab. Publicly available safety data from e.g., literature, and national and European assessment reports are insufficient for sex-specific assessment of side effects. The newly constructed $2FES_2$ visual summary uses data visualization, substantiated by statistical analysis. This aids in the exploration, comparison and comprehensible reporting of multiple side effects at once, detecting sex-specific safety signals. This method could be applied to assess potential sex differences in other medicines.

Abstract

Introduction: Advocacy groups claim that women are underrepresented in medical research, resulting in unsafe treatments. Osteoporosis is a chronic disease that causes bone fragility and risk of fractures and occurs predominantly in women. Denosumab is investigated for osteoporosis treatment separately in men and women, resulting in much sex-specific data. Nevertheless, it remains unclear whether the treatment is equally safe for both sexes. The objective of this research is to evaluate potential sex differences in the safety profile of denosumab. In a broader scope, we aim to formulate a methodology for sex-specific safety analysis, transcending denosumab and becoming applicable to other medicines and subpopulations.

Methods: We compared 4 groups (men and women, each with denosumab and control (placebo) treatments). There is no golden standard for the assessment of differences in safety of a medicine between the sexes, because gender differences are known to occur in control treatment, hampering the interpretation of denosumab data. Adverse event (AE) data from representative studies on denosumab were retrieved from public sources and the Dutch regulatory authority's database. Visual summaries were made for exploration of adverse event profiles, followed by statistical analysis by Fisher's Exact test, meta-analysis and multivariate logistic regression analysis.

Results: Publicly available data lacked detail for these analyses. The regulatory authority database provided workable data from 8558 females and 1927 males for analysis. Analyses detected similar safety signals that concern sex differences in the incidence of serious adverse events under denosumab and control treatment (example: Musculoskeletal disorders, relative risk (RR) of men vs. women in control arm = 0.48 (Fisher's Exact p-value = 0.03), RR in denosumab arm =0.43 (p = 0.013). In none of the system organ classes was serious adverse event incidence affected by both sex and treatment (example: Musculoskeletal, logistic regression: sex [Male] x treatment[denosumab] p =0.824). We constructed a new plot: the2 factor efficacy and safety scatter ($2FES_2$), enabling identification of sex differences, treatment differences and those sex differences that are attributable to denosumab.

Conclusion: Statistical analysis of representative studies does not reveal important sex differences that are attributable to the presence of denosumab. Publicly available safety data from e.g., literature, and national and European assessment reports are insufficient for sex-specific safety assessment. The newly constructed $2FES_2$ uses data visualization, substantiated by logistic regression. This aids in the exploration, comparison and comprehensible reporting of safety profiles, detecting sex-specific safety signals. This method could be applied to assess other safety parameters such as discontinuations, or other medicines.

Table of content

1. Background

1.1 Sexes in clinical trials

Historically, clinical trials have not always adequately included different population groups. Males, frequently of the Caucasian race, were the norm study population, assuming that women and other races would have the same response to medicines. The thalidomide (Softenon) tragedy in the 1960s is an example that demonstrated that drug responses can differ between species. Thalidomide was marketed as a nonaddictive sedative that was very effective in preventing morning sickness in pregnant women. Premarketing trials in female pregnant rats did not reveal adverse reactions, and there was no human research that elucidated any sex-related issues. After thalidomide reached the market, it caused severe birth defects in over 10,000 children. This drastically changed the way drugs are tested nowadays(1).

As amendments were made which required drug manufacturers to prove that a drug is safe and effective through animal experiments and regulated trials in human, it has become clearer that some diseases and treatments have differing effects between, but also within species. Biological differences between males and females can affect the efficacy and toxicity of a medicine, as a result of physiological differences such as body mass and body surface area, but also due to differences in pharmacokinetics (PK) and pharmacodynamics (PD). Drug absorption, distribution, metabolism, PD and adverse events are also influenced by sex hormones(2).

To address the concerns about sex-specific drug efficacy and safety issues, policies have been developed for the inclusion of different population groups (women and minorities). One of these policies is the EU clinical trial regulation No. 536/2014. This regulation calls for more transparency of clinical trials data and adequate inclusion of age and gender groups to represent the population that is likely to use the medicinal product investigated in the clinical trial(3).

This research focusses on sex and medicine safety, specifically how to assess potential sex differences using data retrieved from controlled clinical trials.

1.2 Safety

The safety profile of a medical product concerns the medical risk to the subject, usually assessed in a clinical trial by adverse events (diseases, signs and symptoms) and physical examination, laboratory tests (including clinical chemistry and haematology), vital signs and other special safety tests (e.g. ECGs, ophthalmology)(4).

An adverse event (AE) describes any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment(5).

A serious adverse event (SAE) is an adverse event that results in death, is life-threatening, requires (prolonged) hospitalization, results in an ongoing or significant incapacity or interferes substantially with normal life functions, or causes a congenital anomaly or birth defect. Medical events that put a person in danger or require medical or surgical intervention to prevent one of these outcomes can also be considered an SAE. It is important to note that (serious) adverse events do not imply a causal relationship with the intervention or treatment.

According to the guidelines for the evaluation of safety and tolerability of the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human User (ICH), all safety parameters need attention. We aim to compare the safety profiles of men and women. The difficulty with comparing safety profiles, is that there is no standardized method for the measurement and

reporting. Depending on the study objective, adverse events can be collected and reported spontaneously or through active event monitoring; all emerging events can be screened, or only pre-specified events. Some published papers only report hospitalizations or fatal adverse events, investigational product discontinuations or complete study withdrawals. There are various ways in which these parameters are reported: you can report the incidence of (serious) adverse events or the subject incidence of (serious) adverse events per unit of time; you can report the time to onset of the first event (per person year); sometimes only the events that are deemed related to the investigational product can be reported or only those events that occur in more than 2% of any treatment arm.

Despite these numerous approaches to safety reporting, clinical trials are primarily designed and powered to detect medicine benefits with safety assessments as a secondary outcome. They lack power to detect rare but potentially serious adverse events. The ICH does not have specific guidelines for the clinical investigation of medicinal products in women (or men) as a special population, partly because some form of evaluation for sex-related effect is generally conducted and expected(6). Therefore, there is no defined gold standard on sex-specific safety reporting and (statistical) analysis, which makes the nature of safety data heterogenous and difficult to analyse across trials(7).

In order to learn more about potential sex differences in the safety of medicinal products, the drug denosumab is selected as the subject of this case study. Denosumab was selected not only because there is a lot of study data in women, but also because there are separate trials in men and women. We expect that this results in a large availability of sex-specific data. We will be looking for sex-specific safety signals. A safety signal is information on a new or known adverse event that may be caused by a medicine and requires further investigation.

1.3 Osteoporosis and denosumab

Osteoporosis is a medical condition in which the concerns about gender-bias seem to be reversed. It is considered a women's disorder, as female sex and age are associated with osteoporosis incidence. In the Netherlands, over 500.000 people were diagnosed with osteoporosis in 2019, of which 426.100 (84%) were female(8). The prevalence ratio (426.100 / 507.260 = 0.84) translates to the clinical studies, where a major part of studies is conducted in women.

Osteoporosis is a chronic and progressive skeletal disorder characterized by an increase in bone remodelling, decreased bone mass, microarchitectural deterioration and compromised bone strength. The resultant bone fragility predisposes people with osteoporosis to an increased the risk of fracture(9). The disorder can occur as a primary or secondary condition. Primary osteoporosis is mostly due to age-related loss of bone, but there are other risks factors such as race, dietary calcium deficiency, sedentary lifestyle, alcohol use, family history and cigarette smoking. Secondary osteoporosis results from the presence of other diseases and/or medications such as hormone ablation therapy or being underweight(8).

1.3.1 Primary osteoporosis

Post-menopausal osteoporosis is the main form of primary osteoporosis, as age and menopause are the two main determinants of osteoporosis and fracture risk. Fractures are more common in women, in part, because of lower bone mass and also because there is less competing mortality (women tend to live longer than men)(10). Men can suffer from primary osteoporosis as well, but male osteoporosis is less prevalent. According to predictions based on demographic development of the Dutch national institute of health and the environment (RIVM), the prevalence of osteoporosis will

increase from 2005 until 2025. The total amount of women aged ≥ 55 with osteoporosis is predicted to increase with 37%, compared to an even larger increase of 50% in the male population(11).

1.3.3 Glucocorticoid-induced osteoporosis

Corticosteroids are hormones produced into adrenal cortex with an anti-inflammatory and immunosuppressive effect. Natural and synthetic steroid medication exists and can be given locally as injections, eye or eardrops and skin creams, or systemically through oral medicines or intravenous or intramuscular injection. Glucocorticoids (GCs) are a subgroup of corticosteroids. Prednisone, a systemic GC, can be prescribed to treat a spectrum of conditions such as rheumatoid arthritis, vasculitis, chronic obstructive pulmonary disease (COPD), sarcoidosis and inflammatory bowel disease. Inflammation itself, but also glucocorticoids are associated with low bone mineral density and drive bone loss(12, 13). A study in healthy volunteers showed that prednisone increases the risk of fracture by decreasing the rate of bone formation and by decreasing the number and activity of osteoblasts and osteoclasts (fig. 1). The changes are reversed after prednisone discontinuation(14). Glucocorticoid-induced osteoporosis (GIOP) is the most common form of secondary osteoporosis.

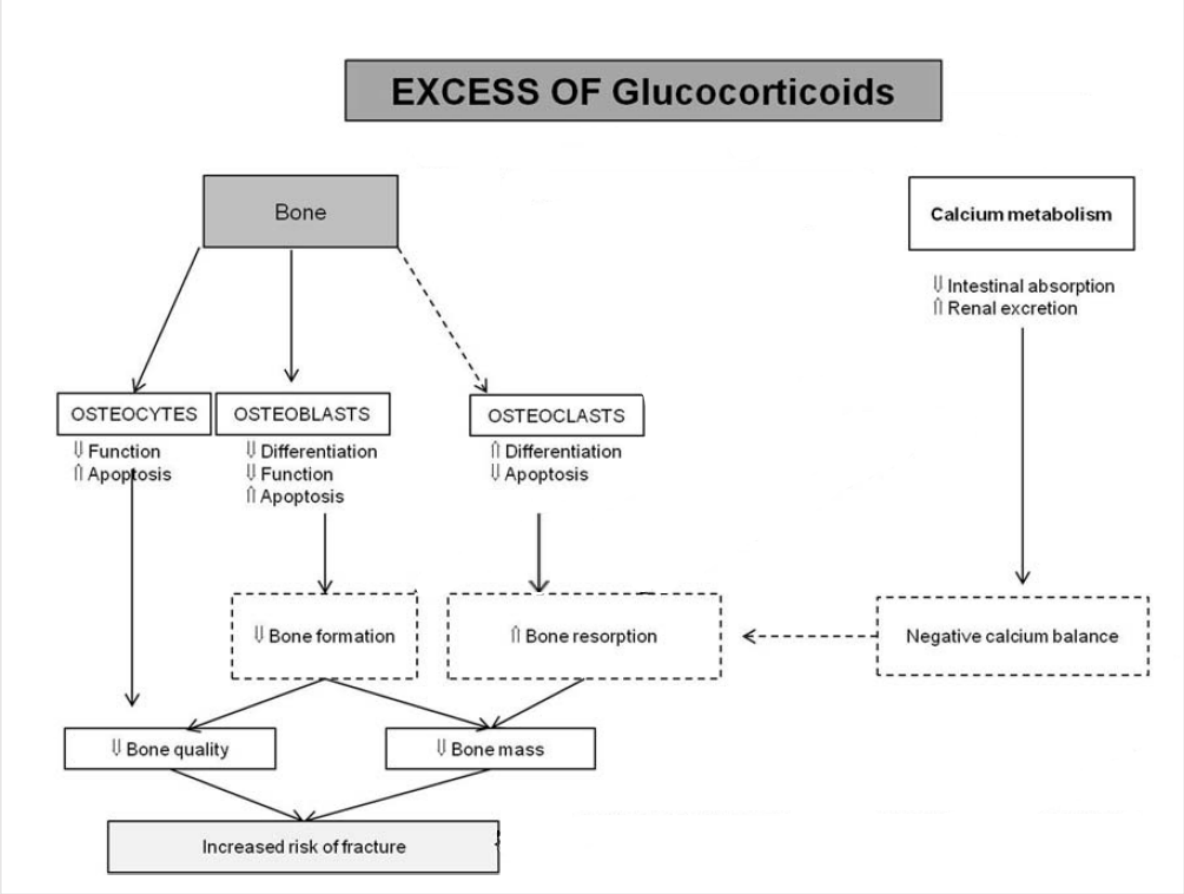

Figure 1. Pathophysiology of glucocorticoid-induced osteoporosis. Adapted from Briot et al. 2015(15).

1.3.2 Cancer treatment-induced bone loss

Endocrine therapy is a medical treatment that adds, blocks or removes hormones. Androgen deprivation therapy and aromatase inhibitor therapy are treatments given to respectively men with nonmetastatic prostate cancer and women with hormone receptor positive early-stage breast cancer. It is known that these treatments compromise bone health. Aromatase inhibitors suppress the conversion of androgens to oestrogens, resulting in oestrogen depletion, which leads to lower

BMD(16). Androgen-deprivation therapy increases bone resorption, thereby reducing BMD and increasing the risk of fracture(17).

1.3.3. Denosumab

If the balance between osteoclast (cells that degrade bone to initiate bone remodelling) and osteoblast (bone forming cells) activity is unbalanced, pathological conditions such as osteoporosis can occur. Denosumab is a biological agent prescribed for the treatment of osteoporosis. It is a human monoclonal IgG2 antibody, that inhibits osteoclast activity by targeting and temporarily preventing the interaction of the RANK receptor and its activator: receptor activator of nuclear factor-κβ ligand (RANKL), present on osteoclasts. RANKL is a cytokine that plays a pivotal role in the formation, function and survival of osteoclasts(8, 14) (fig. 2).

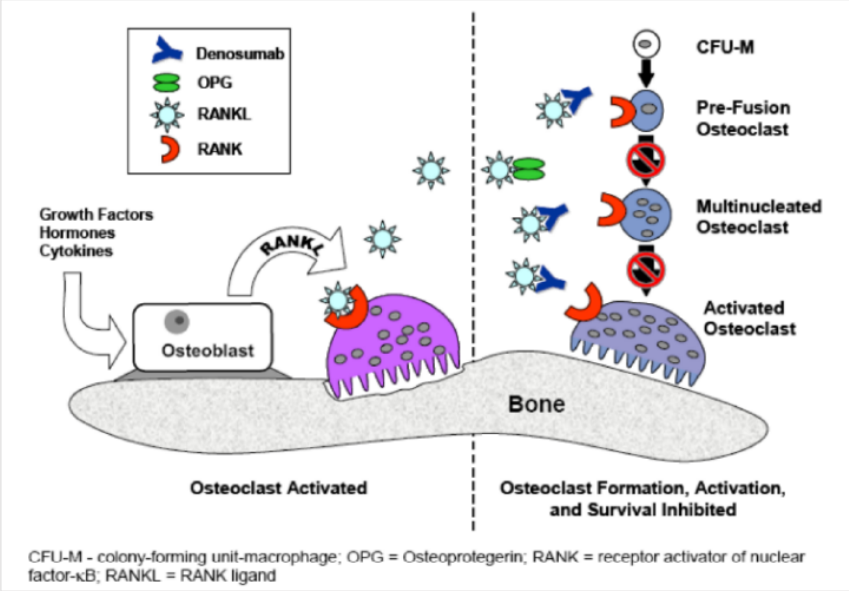

Figure 2. Mechanism of action of denosumab (source: European Public Assessment Report Prolia 2010)(18).

By inhibiting osteoclast activity, denosumab reversibly inhibits bone resorption, which results in reduced bone turnover followed by increased bone density and decreased fracture risk(8). As RANKL production is upregulated during glucocorticoid use, denosumab has been studied as a pharmacological treatment to counteract GIOP. A phase 2 trial with rheumatoid arthritis patients who were taking glucocorticoids showed that biannual injection of 60mg and 180mg denosumab effectively improved bone mineral density (BMD) and suppressed bone turnover after a study period of 12 months(19).

Denosumab has also been researched in patients with metastatic cancer such as castration resistant prostate cancer, multiple myeloma and non-small-cell lung cancer. Patients frequently experience osteoclast-mediated bone destruction, resulting in complications such as fractures, spinal cord compression or hypercalcemia, collectively known as skeletal-related events (SREs). Several studies have been conducted and show effectiveness of denosumab to prevent or delay SREs and bone metastases(20, 21). Denosumab was approved in the U.S.A. for the prevention of skeletal-related events in patients with solid tumours and bone metastases(22).

Denosumab has been authorised since 2010 with an indication to treat osteoporosis in postmenopausal women (primary osteoporosis), and bone loss associated with hormone ablation therapy in men with prostate cancer (secondary osteoporosis). A randomized trial by Saag et al. 2018(23) assessed the efficacy and safety of denosumab against an active comparator treatment in

GIOP, leading to the additional indication for treatment of secondary osteoporosis associated with sustained systemic glucocorticoid therapy and to treat people who are starting or have recently started long-term glucocorticoid therapy.

1.4 Research question

Advocacy groups claim that there is an underrepresentation of women medical research, resulting in unsafe treatments. In denosumab research, the gender bias seems to be reversed, as most research has been conducted in women, potentially resulting in unsafe treatment for men. Osteoporosis is a chronic disease that causes bone fragility and risk of fractures and occurs predominantly in women. Denosumab is a monoclonal antibody with an anti-resorptive effect, used to increase the bone mineral density and decrease the risk of fracture in osteoporotic patients. Many studies investigate denosumab separately in osteoporotic men and women. Nevertheless, it remains unclear whether the treatment is equally safe for both sexes.

There is no defined gold standard or guideline on sex-specific safety analysis, despite gender questions that keep arising in society. We use denosumab as a case study, because we expect a large availability of sex-specific safety data. The objective of this research is to evaluate potential sex differences in the safety profile of denosumab, by searching for sex-specific safety signals. In a broader scope, we aim to formulate a methodology for sex-specific safety analysis, transcending denosumab and becoming applicable to other medicines and subpopulations.

2. Methods

There is no golden standard on how to assess differences in safety of a medicine between the sexes because gender differences are known to occur in control treatment hampering the interpretation of denosumab data. We therefore developed our own approach. Our approach consists of:

1. Identifying denosumab safety data sources
2. Selecting a safety parameter for analysis
3. Identifying data visualization methods
4. Identifying statistical analysis methods
5. Applying and ranking the identified methods

Firstly, we undertook an online search of clinical trial registries, published scientific literature databases and regulatory authority platforms, to identify their suitability on providing sex-stratified denosumab safety data. From these sources, controlled clinical trials were selected to form our safety analysis set. Secondly, we decided to select a single safety parameter for safety assessment. Thirdly, we undertook an online search to identify literature describing methods for safety analysis, including applicable data visualization methods and statistical methods for adverse events. The selected visualization and statistical methods were applied to our safety analysis set. Methods were tested on their ability to detect sex-specific safety signals, thereby generating a hypothesis on possible sex differences in the safety profile of denosumab. The same ranking system is used previous reviews(24) was applied, which we expanded with aesthetics as an additional ranking criterium.

Ranking criteria

1. *Ease of implementation*
2. *Ease of comprehension*
3. *aesthetics*

Visual summaries were made for initial exploration of adverse event profiles. Subsequent statistical analysis was performed to confirm any subjectively identified safety signals concerning a treatment effect, a sex effect or a sex effect attributed to denosumab treatment. All data visualization and statistical methods were performed in R: A Language and Environment for Statistical Computing(25, 26).

2.1 Identification of data sources

An online search was conducted to identify sources that provide denosumab safety data. We chose to only use controlled data, because uncontrolled data, open label studies, single arm studies, observational data, spontaneous safety signal collection and post marketing data have some major difficulties. Non-randomized and uncontrolled/single arm trials tend to measure different incidences of adverse events depending on the preferred term(27); background event rates in the post-marketing rate may differ, as the treatment indication can be different/broader that the inclusion criteria for trials (confounding)(28); spontaneous reporting is known to suffer from underreporting up to 95%, depending on the severity and novelty of the event, the drug class and the organ system class(29).

The clinical trials registry clinicaltrials.gov was used to filter and select randomized controlled clinical trials which compiled would be representative for the real-world osteoporotic population in terms of sex prevalence and types of osteoporosis (primary osteoporosis; post-menopausal and primary in men, secondary; caused by glucocorticoid therapy, hormone ablation therapy for prostate cancer and aromatase inhibitor therapy for breast cancer). Selection criteria were randomized, controlled clinical trial, study completed with results, adult patient population ($\geq$ 18 years old), published and openly available full text article, trial duration $\geq$ 1 year, at least 100 participants per treatment arm, a dosing regimen of subcutaneous injection of 60mg denosumab once every 6 months, osteoporotic patient population, available safety data with known sex stratification if necessary and compatible study objectives. Adverse event data from representative studies on denosumab were retrieved from public sources and were supplemented with data from the Dutch regulatory authority's database if publicly available data were incomplete.

Usually, all participants who received at least 1 dose of investigational product (IP) are included in the safety analysis set. Therefore, we used the same subjects in our safety analysis set of the selected trials.

2.2 Events

To conduct sex-specific safety analysis we selected serious adverse events as a parameter that represents safety, as this is deemed the least subjective to reporting bias from the patient, but still covering all possible events (including fatal events). We compared the subject incidence of serious adverse events in 4 groups (men and women, each with denosumab and control (placebo) treatments) at the system organ class level. We selected the system organ class level because it has twenty-six manageable classes we can use for analysis. This is the broadest level the MedDRA (Medical Dictionary for Regulatory Activities) hierarchy. Safety signals found at this level could be explored further into more specific levels later.

We distinguish between parameters that describe safety or tolerability of a medicine. Tolerability represents the degree to which overt adverse reactions can be tolerated by the subject(4). A therapy is deemed tolerable if subjects are willing to take and adhere to a drug or therapy. Willingness deviates from subject to subject. It depends on how much the disease affects the subject's quality of life and how much relieve (and potential side-effects) the therapy offers (benefit-risk balance). We

define discontinuation of the investigational product and complete withdrawal from a clinical trial as parameters of tolerability.

The relatedness of (serious) adverse events is reported dichotomously by the investigator with a yes or no response to the question: Is there a reasonable possibility that the event may have been caused by the IP? For the purpose of this study, this is considered a subjective observation and no further inquiries into relatedness will be made.

2.3 Data visualization methods

The utility of graphical methods was investigated for qualitative analysis of the sex-specific safety profile within and between treatment groups. We examined the ability of data visualization methods to identify important safety signals between sexes. Subsequently, quantitative analysis of the adverse event data was performed through various statistical methods to comment on the effectivity of the graphical data presentation methods.

Visualization methods should aid in the exploration of the full safety profile to generate sex-specific safety signals of events that need deeper evaluation. We searched for methods that classify as: for presentation of binary data; for use on multiple emerging (not prespecified) events. Based on literature, the following visual summary methods for displaying adverse event data were considered promising: the bar chart, dot plot, heatmap, forest plot and the volcano plot. Methods were applied to denosumab safety data, discussed and ranked.

We aimed to use these graphical methods to evaluate and enhance the presentation of sex-specific safety data. The performance of these methods depends on the objective of the study. To examine the performance in our research, we developed a ranking system based on subjective criteria we deemed important. Each method is relatively ranked based on their ability to detect sex-specific safety signals, the ease of implementation, ease of comprehension and the aesthetics.

2.4 Statistical analysis methods

Statistical methods should be able to use the same data as used in the data visualization methods. We searched for methods that can analyse discrete, categorical data (involving two possible outcomes: experiencing an adverse event vs. no adverse event, as opposed to continues data, such as weight or plasma calcium levels). Methods were selected if they could perform analysis on the total safety profile, using summarized safety data (as patient-level data was not available). Analyses are performed post-hoc, therefore methods that are used for prespecified adverse event data are excluded. Based on literature, the following statistical methods for analysis of adverse event data were considered promising: Fisher's exact test, meta-analysis, and multivariate logistic regression found in a review by Phillips et al 2020(24). Methods were applied to denosumab safety data and subjectively ranked based on their ability to detect sex-specific safety signals, the ease of implementation and ease of comprehension.

3. Results

3.1 Suitability of data sources

Clinical trials registries, and public assessment reports were used to filter the clinical trials that would form our safety analysis set (fig. 3). Published papers, clinical trial registries, public documents from regulatory authorities (EMA and FDA) and not-publicly available regulatory submissions of the selected trials were searched assessed. All sources presented safety information to a certain extent. Often only events that were of interest were reported, or only those events that occur in more than 2% of any treatment arm. This differs per published paper and was therefore not compatible across trials. The summarized subject incidence of serious adverse events in the first year on trial was retrieved from clinical study reports submitted to the Dutch regulatory authority database (ICI) and used for graphical presentation and subsequent statistical analysis. The summarized data contained information about the study, treatment arm, group size, system organ class and sex. The use of grey literature was necessary, as public sources did not provide data that was in depth and extensive enough(30). Grey literature: all types of research materials that have not been made available through conventional publication formats. More information about the suitability of different identified data sources can be found in Appendix A.

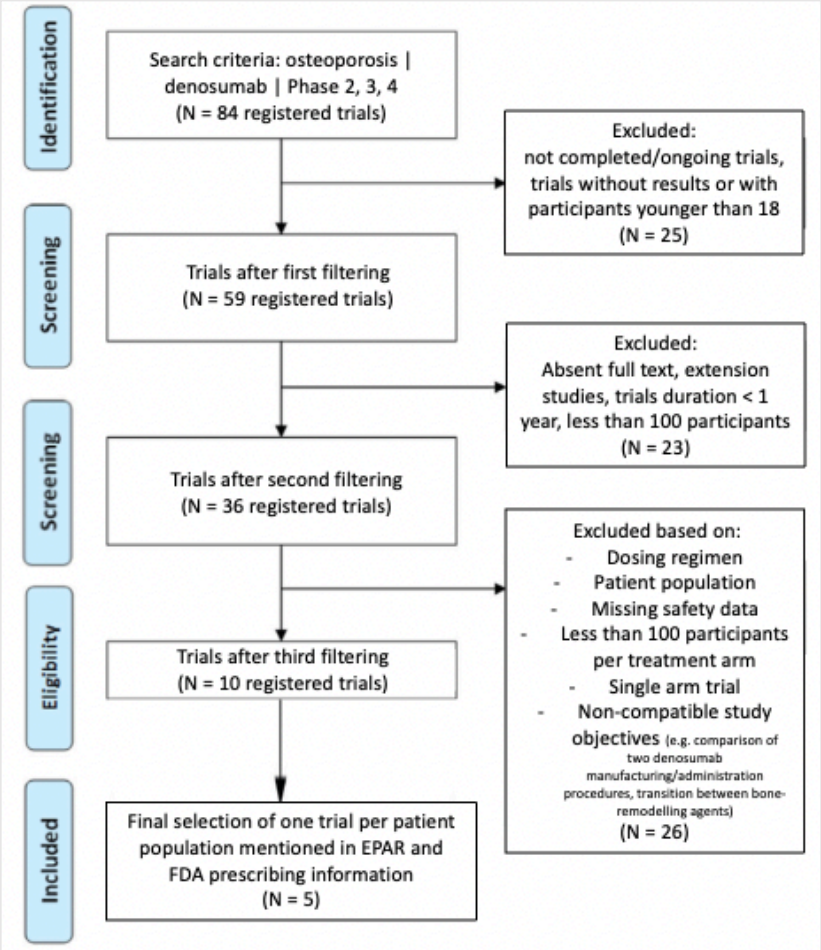

Figure 3. Flow chart of trial selection process for safety analysis set formation. Selected trials are: Cummings et al. 2009(14), Smith et al. 2009(17), Orwoll et al. 2012(31), Gnant et al. 2015(16) and Saag et al. 2018(32).

The five selected trials form a compiled safety analysis set that present every indication for denosumab treatment defined by the European and U.S. drug regulatory authority (table 1). The safety analysis set included 10.485 subjects, of which 1.936 (18%) were men. This results in a participation to prevalence ratio of 0.97 (PPR = 0.97). A ratio between 0.8 and 1.2 is considered as proportional sex-representation, compared to the real-world sex distribution in the osteoporotic population(33-35). A total of 1.247 subjects suffered a serious adverse event, of which 283 (23%) were men.

Table 1. Characteristics of selected randomized controlled clinical trials.Q6M = Single dose at the first and sixth month of each year, SQ = subcutaneous.

Study number	Published article	Population	Intervention	Comparator	Sex distribution	
					Men	women
1	Cummings 2009(14)	Post-menopausal	Q6M SQ injection of 60mg denosumab	6QM SQ injection of placebo	0 (0%)	7762 (100%)
2	Smith 2009(17)	bone loss associated with androgen deprivation therapy for prostate cancer)	Q6M SQ injection of 60mg denosumab	6QM SQ injection of placebo	1456 (100%)	0 (0%)
3	Orwoll 2012(31)	Primary osteoporosis	Q6M SQ injection of 60mg denosumab	6QM SQ injection of placebo	240 (100%)	0 (0%)
4	Gnant 2015(16)	Osteoporosis associated with aromatase inhibitor therapy for breast cancer	Q6M SQ injection of 60mg denosumab	6QM SQ injection of placebo	0 (0%)	249 (100%)
5	Saag 2018(32)	Glucocorticoid-induced osteoporosis (GIOP)	Q6M SQ injection of 60mg denosumab	6QM SQ injection of placebo	231 (30%)	547 (70%)

3.2 Strategies

Literature on data visualization methods for presentation of adverse event data and statistical analysis methods to compare treatment arms, sexes and to perform inter-trial comparisons were searched via PubMed and Google Scholar. Both theoretical as applied methods were analysed to see if they can be incorporated into our methodology to assess potential sex differences in safety data from randomized controlled clinical trials. A review by Phillips et al. 2020(24) summarizes a selection of (statistical) methods for the analysis of adverse event data.

3.2.1 Data visualization

Adverse event reporting often defaults to simple frequency tables, whilst there is a wide range of data visualization methods. Other methods could possibly present data more clearly or could be easier to read and faster in detecting safety signals from events in which the incidence is subject to sex.

Next to the most frequently used method: summarizing subject incidences of events in a contingency table, five other methods met our classification criteria. These were applied to our formed safety analysis set and criticized based on their ease of implementation, easy of comprehension and aesthetics (table 2). The ranking of the methods as well as the system organ classes from which potential sex-specific safety signals arose are subjective to the viewer's programming capabilities and opinion, as well as the objective of research conducted. For discussion on each visualization method, see appendix B.

All visualization methods implicate that the incidence of serious adverse events in the system organ class "Injury, poisoning and procedural complications" is potentially subject to sex differences. Three methods detect a safety signal from system organ class (SOC) "Respiratory, thoracic and mediastinal disorders" and "Gastrointestinal disorders" one detects "Cardiac disorders", one detects "Renal disorders" and another one detects "Vascular disorders". All these detections are subjective to the viewer, except from the forest plot and volcano plot, in which safety signal detection is based on the calculated risk ratio and confidence intervals and the p-value of Fisher's exact test.

Table 2. Ranking of the visualization methods based on ease of implementation, ease of comprehension and aesthetics. Appendix B presents imaging and discussion of each method separately. SOCs = System Organ Classes, names are abbreviated. Full names of the system organ classes can be found in Appendix C1.

Rank	Method	Depicted safety information	Main advantage	Main disadvantage	Detected sex-specific safety signals: SOCs (Subject to viewer)
1	Volcano plot (Appendix B1)	Relative risk + Fisher's exact p-value	Detects SOCs based on p-value	No direct comparison between sexes	Injury
2	Forest plot (Appendix B2)	relative risk per sex + CI + (weighted) average relative	Detects SOCs based on CI	Overly large plot needed to display all SOCs Statistical analysis	Injury
3	Dot plot (Appendix B3)	Absolute subject incidence + relative risk + CI	Detects SOCs based on CI	Interpretation requires two plots	Gastrointestinal Injury Respiratory
4	Contingency table	Absolute subject incidence numbers and percentage) per	Straightforward presentation of all SAE counts	Only presents numbers Only presentation, no	Injury Cardiac Gastrointestinal Respiratory
5	Heatmap (Appendix B4)	Relative SOC incidence (%) per sex per treatment arm	Easy to spot major outliers	Relative incidence is misleading for interpretation Only	Respiratory Renal Injury Gastrointestinal
6	Stacked bar chart (Appendix B5)	Relative SOC incidence (%) per sex per treatment arm	Easy to spot major outliers	No common base Presentation of relative incidence is misleading	Gastrointestinal Injury Vascular respiratory

We gave the stacked bar chart the lowest ranking. The method was easy to implement, because the original data did not need heavy conversion before visualization. It is an aesthetically pleasing plot, because there are only four bars that represent the four groups to compare (male and female, divided in control and denosumab treatment) and within each bar the system organ classes are

represented by different colours. However, the stacked bar chart depicts the relative abundance of each SOC within each group. Because SOCs lack a common base, comparison of SOCs between different groups, and therefore detection of sex-specific safety signals, is not very exact. Comparison could even be misleading, because all bars are evenly tall (100%), while the subject count of each group differ greatly (female denosumab treatment arm = 4191 vs. male denosumab treatment arm = 969). The volcano plot was ranked in first place (fig. 4). Implementation is more difficult due to the statistical analysis data results that are used for plotting the SOCs. The ease of comprehension and aesthetics are ranked high and pleasing. Because the position of the SOC on the x-axis shows if the serious adverse incidence is increased, decreased or similar between control and denosumab treatment and the position on the y-axis represents the significance of the relative risk. This makes it easy to detects sex-specific safety signals from SOCs that have a significant difference.

A disadvantage that all visualization methods have is that they do not allow comparison between both control and denosumab while simultaneously comparing the incidences in men and women. It is either visualization of control vs. denosumab treatment, separately for men and women, or men vs. women, separately per treatment arm.

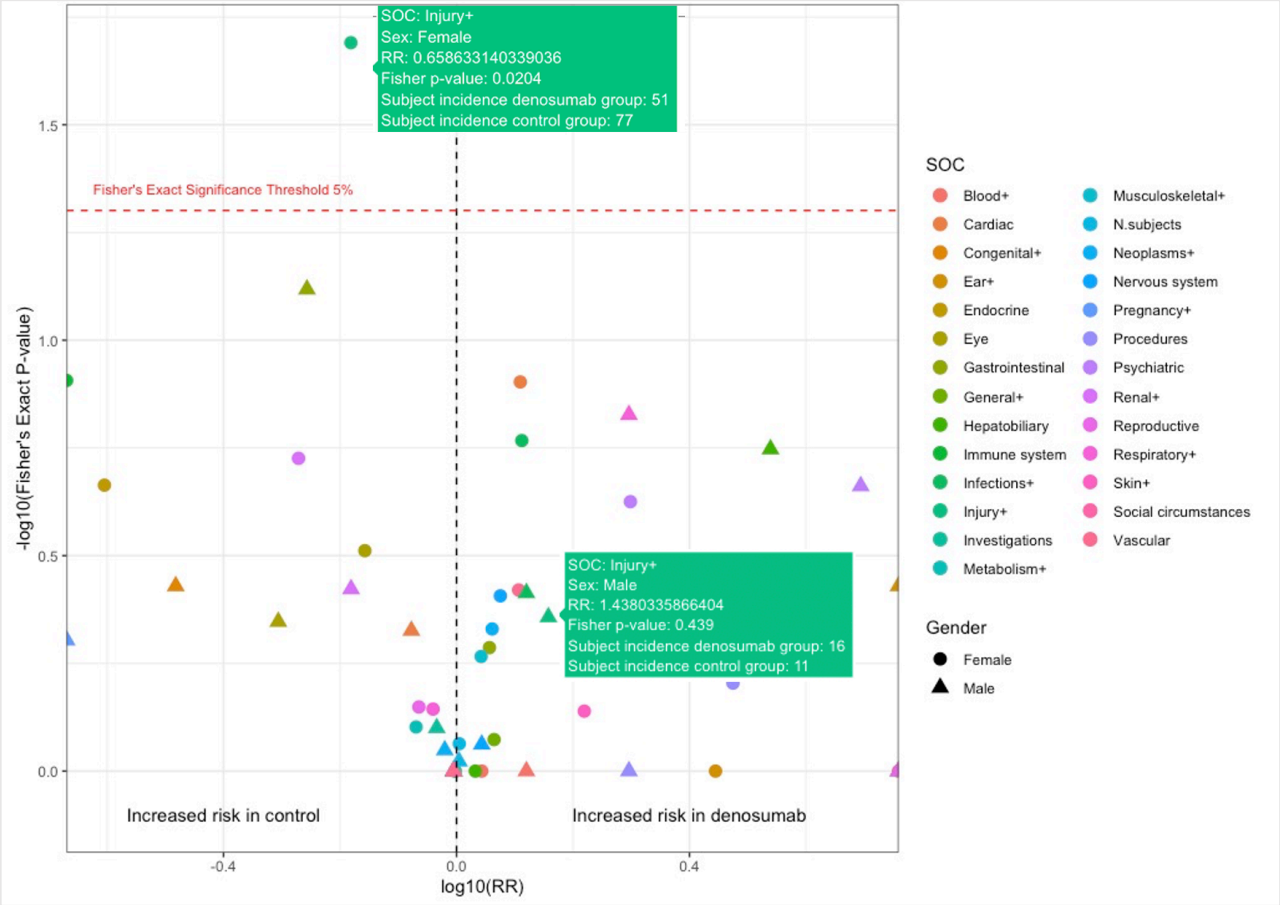

Figure 4. Volcano plot displaying the risk ratio of control vs. denosumab treatment for men and women separately, combined with statistical significance of the risk ratio determined by Fisher's exact test. SOC = System Organ Class, names are abbreviated. Full names of the system organ classes can be found in Appendix C1.

3.2.2 Statistical analysis

Statistical methods for the analysis of safety data in randomised controlled trials (RCTs) are rarely used, and there is a reliance on simple display methods such as the contingency table to relay safety information.

Three statistical approaches are selected from the review by Phillips et al(24): Fisher's exact test, meta-analysis and multivariate logistic regression analysis. The selected methods were classified as: for analysis of binary data; for use on emerging (not prespecified) events; for final/one analysis; for analysis of multiple events, to apply to our formed safety analysis set. These strategies were also ranked based on their ease of implementation and comprehension (table 3).

Table 3. Ranking of statistical analysis strategies based on ease of implementation and ease of comprehensions. Appendix C presents the results and discussion of each method separately. CI = confidence interval, SOCs = System Organ Classes, names are abbreviated. Full names of the system organ classes can be found in Appendix C1.

Rank	Method	Depicted safety information	Main advantage	Main disadvantage	Detected sex-specific safety signals: SOCs
1	Multivariate logistic regression analysis (Appendix C1)	Sex and treatment effects and their interaction	Tests for sex-effect and treatment-effect and p-value	Does not assess all SOCs in one go	No interactions between sex and treatment effect
					Treatment effect on Injury and Ear
					Sex effect on total SAE incidence, Cardiac, Musculoskeletal, Neoplasms, Gastrointestinal, Vascular, Renal, Metabolism, Investigations, Congenital, Pregnancy
2	Meta-analysis (Appendix C2)	SAE risk ratios and confidence intervals per sex	Detects SOCs based on CI's and shows direction of association	Cannot test for treatment and sex effects simultaneously	Injury (in women)
3	Fisher's exact test (Appendix C3)	Association of treatment and SAE incidence per sex	Easy to detect safety signals from SOCs based on p-value	Direction of association is not known immediately	Injury (in women)

Fisher's exact test assesses if there is a non-random association between sex and serious adverse event incidence. Implementation is easy, but laborious because the test for each SOC must be coded separately. The outcome is a relative risk of control vs. denosumab treatment and a p-value, which is easy to understand. A p-value below 0.05 (or another threshold that you pick) is deemed significant.

Meta-analysis got ranked higher, because the test is less laborious, all data can be put into a single analysis and does not require any conversion, whilst giving the same or even more information

compared to Fisher's exact test. The results of the test are the relative SAE risk of control vs. denosumab treatment per system organ class, per sex. This is shown in a forest plot. A 95% confidence interval is calculated and depicted as well. A CI that does not include the value 1 suggest that the relative risk is significant, making it easy to detect SOCs with sex-differences. The random-effects model is also automatically implemented. This calculates an over-all, average relative risk for all SOCs. However, the calculation is incorrect, because it adds up the subject counts of every SOC, while this is the same population. This causes the random effects model to use a SAE subject count and population size that is larger than it is.

Multivariate logistic regression analysis is ranked highest. Implementation works like that of Fisher's exact, as it is laborious. The output is extensive and easy to comprehend. It gives the odds ratios with p-values compared to given base condition. In this case the base condition is being of female sex and being in the control treatment arm. For every change: male sex, denosumab treatment or both, and odds ratio and p-value is given. This answers if there is a significant sex-effect, a treatment-effect and or an interaction. A significant interaction implies that a found sex-difference in a SOC is specific for the denosumab treatment. For more discussion on each statistical analysis method, see appendix C.

3.2.3 Two-factor efficacy and safety scatter ($2FES_2$)

To improve sex specific safety reporting and safety signal detection, we combined visual summarizing with regression analysis into a new plot. We use this plot for the reporting of both sex and treatment effect on safety, but it is possible to also incorporate efficacy parameters in this plot. We therefore named it 2-factor efficacy and safety scatter (2FES2). The x-axis displays the significance of the sex effect on SAE incidence, while the y-axis displays the treatment effect on SAE incidence. Significance is determined by multivariate regression analysis. SOCs above the horizontal threshold are detected with significant effect of [male sex], SOCs to the right of the vertical threshold are detected with a significant effect of [denosumab treatment]. SOCs from which both treatment and sex-effects are detected, would be plotted in the orange highlighted area (fig. 5).

The $2FES_2$ plot shows an empty orange area, indicating that [denosumab treatment] and [male sex] combined do not cause a significant change in the incidence of serious adverse events. This means that the interaction between these two factors is not an accurate predictor. In conclusion, our safety dataset suggests that there are no sex differences in the safety profile of denosumab.

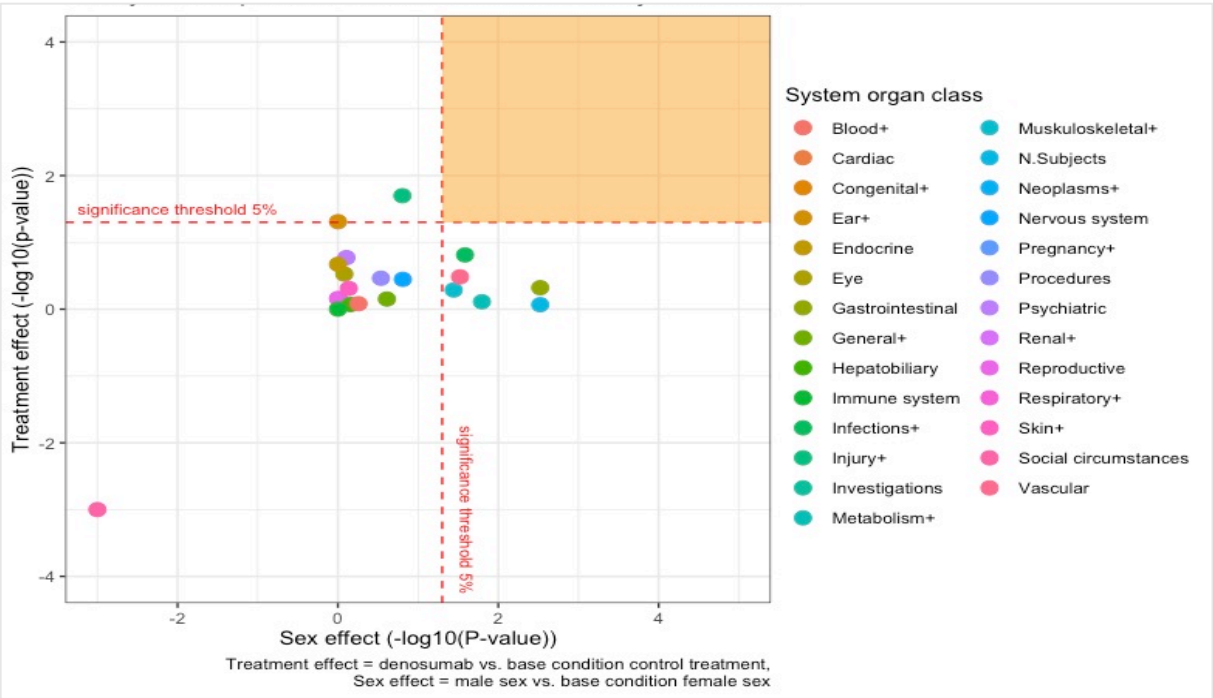

Figure 5. Two-factor efficacy and safety scatter plot of serious adverse event incidence. factors are sex effect and treatment effect, determined by multivariate logistic regression analysis for each system organ class. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

3.3 Denosumab safety

Subjective assessment of the data visualization methods revealed potential sex-specific safety signals in the system organ classes gastro-intestinal disorder, Injury, poisoning and procedural complications, vascular disorders, respiratory, thoracic and mediastinal disorders, renal and urinary disorders, ear and labyrinth disorders and cardiac disorders. These SOC's are detected because the visual summaries seem to display a control-denosumab incidence ratio that differs between the male and female subset of pooled study population. Thresholds for the final detection of safety signals are decided upon by the investigator. In case of the forest plot and the volcano plot the threshold is statistically determined and therefore more objective and reproducible. We therefore observe the strongest sex-specific safety signal to originate in the injury, poisoning and procedural complications system organ class.

All three subsequent statistical analyses confirm a treatment effect on the SAE incidence in system organ class "Injury, poisoning and procedural complications". Fisher's exact test and meta-analysis perform analysis separately for men and women. These methods only confirm a treatment effect on SAE incidence in the female subset of the study population. Additional comparisons of men against women, separately for the control and denosumab treatment arm, were also performed. Many safety signals arose, indicating a potential sex-effect on the SAE incidence in both control and denosumab treatment arms. Regression analysis, in which treatment effect and sex-effects are assessed simultaneously, detect these same safety signals. Using regression analysis as the strongest sex-specific safety analysis method, we detected safety signals for indicating that treatment potentially affects SAE incidence in the system organ classes injury, poisoning and procedural complication, and ear and labyrinth disorders. Secondly, we detected safety signals for sex differences in many other system organ classes. No safety signals were detected that indicated potential sex differences in serious adverse event incidence that are attributable to denosumab.

4. Conclusion

There are many ways to assess safety data from randomized controlled clinical trials. In this study we self-generated a methodology that selects representative clinical trials, visualizes serious adverse event data per sex as an initial exploration of the safety profile and statistically assesses sex differences in the subject incidence. This method aims to detect sex-specific safety signals from system organ classes. In the future, this should aid in the decision-making process of benefit-risk management and signal if this is possibly different for each sex.

For each step, multiple ways of visualization and subsequent statistical analysis to perform sex-specific safety of the drug denosumab were assessed. This method generated a safety analysis set of 10.485 subjects representing various osteoporotic sub-populations and both the men and women treated with denosumab.

All visualization methods implicate that the incidence of serious adverse events in various system organ classes is subject to sex differences and treatment differences. The benchmarks used to detect safety signals are subjective to the viewer, except from the forest plot and volcano plot, in which detection is based on the calculated risk ratio and confidence intervals and the p-value of Fisher's exact test. Based on our ranking criteria and objectiveness of safety signal detection, the volcano plot has the most value of the assessed data visualization methods.

Subsequent statistical analyses mostly confirm the subjective findings from the visualization methods and detect even more safety signals, suggesting sex and treatment differences. The regression analysis is most efficient, in giving the most information in the most straightforward way. To visualize these findings for more efficient reporting of sex-specific safety, we generated a new visualization method that integrates the regression analysis results in a scatter plot and named this the $2FES_2$ (2-factor efficacy and safety scatter). The $2FES_2$ shows that our safety analysis set does not detect any events at the system organ class level with a combination of sex- and treatment differences in serious adverse event incidence. Therefore, we do not report any sex-differences in the safety profile of denosumab.

Other conclusions we can draw through the generation of this methodology for sex-specific safety analysis are:

- publicly available data is too heterogeneous and not extensive enough for analysis of the full safety profile per sex.
- Many trials are not suitable or comparable due to small study population sizes, different study designs and different dosing schemes.
- None of the visual summaries and statistical analysis strategies allow efficient analysis of all safety data at once. Multiplicity issues arise when all SOC data is combined (e.g., random effects model of meta-analysis).
- Visualization and statistical methods can use the same data input but vary in presentation and how well they compare the complete safety profiles of both sexes. However, subjectively, safety signals are detected in overlapping system organ classes, so similar conclusions could be drawn from each data visualization method.
- Of the applied strategies, regression analysis displayed in a volcano plot (2 factor efficacy and safety scatter) is most suitable to gain insight in the complete safety profile and for detection of events with sex differences, attributable to denosumab treatment.

4.1 recommendations for the future

For future research that will build upon the knowledge and methodology that was formed through this research, we have several suggestions:

- Use other parameters for sex-specific safety profile analysis (non-serious events, discontinuations of investigational product, longer trial durations (+2 years) or primary vs. secondary osteoporosis).
- Study potential mechanisms that cause sex- and/or treatment effects on adverse event incidence (determine if there is a causal relationship between the drug and the event).
- Study potential sex differences in the tendency to report adverse events and how they affect incidence rates and ratios.
- Investigate the use of the $2FES_2$ plot in assessing potential safety differences between races, ethnicities or age groups (and incorporate efficacy parameters).
- Assess the use of the remaining taxa of adverse event analysis methods described in the review by Phillips et al. 2020 (6): estimation methods (quantification of distributional differences) and decision-making probabilities (methods under Bayesian paradigm) (fig. 6).
- Define a routine method that can be performed to elucidate potential sex-differences.
- Assess how to determine cut-off points for the level of difference and when this requires re-assessment of the benefit-risk balance.

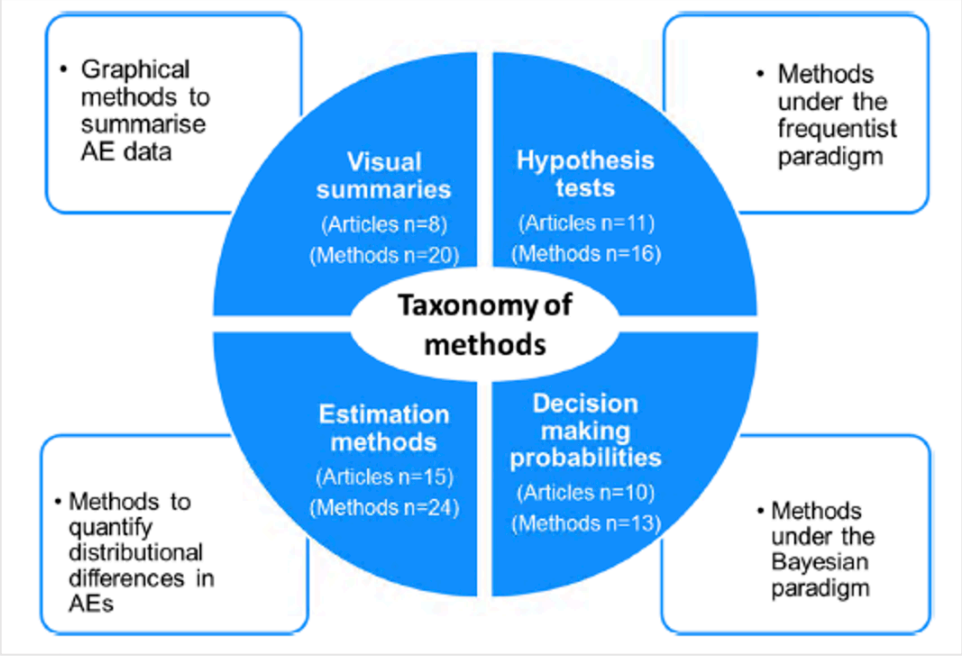

Figure 6. Taxonomy of methods for adverse event analysis(24).

5. Discussion

To heed the calls from society to make medical research more sex and gender aware, in this study we set out to assess potential sex-specific safety issues in the drug denosumab. Our broader aim was to generate a routine methodology to perform sex-specific safety analysis. We found that despite the lack of use, there are many different approaches to report sex-specific safety data analysis.

5.1 Strengths and limitations

The most important strength of this study is the application of multiple visualization and statistical strategies and the generation of the 2-factor efficacy and safety scatter plot. This allows colleagues to select methods that are most suitable for their research based on applied examples.

The most important limitation is considering the subjectiveness of the ranking system, it depends on how well versed the researcher is in coding and statistics, as well as what purpose the visualization and statistical strategies serve. This study only had one researcher that applied and criticized the strategies. Another limitation is the reduction of safety data to binary counts of serious adverse events, while a complete safety profile is broader than that. Safety data is often accompanied by information on occurrences, severity, timing and duration. These factors are important to consider when deciding on the importance of observed safety signals.

A single centre observational study observed a higher prevalence of secondary osteoporosis in men (66.67%) than in women (20.83%) among their observed population(34). If the type of osteoporosis affects the safety or tolerability of denosumab, this could be a possible confounder in sex-specific safety analysis.

5.2 Obstacles

- Information in the clinical study report is often repeated in different tables with slightly different reporting criteria. For example, the subject incidence of treatment-emergent serious adverse events, SAEs of interest, SAEs considered related to investigational product, SAEs reported in $\geq$ 3% (or 5%, or 0.5%) of subjects in either treatment group. This creates unnecessary noise when compiling data from multiple trials, potentially causing incorrect comparison of data that are subject to different reporting criteria.
- Discontinuations of investigational product are not depicted by gender in the denosumab dossier (Saag et al. 2018, GIOP trial)(32). Therefore, all narratives had to be read and counted to discover the male/female distribution of IP discontinuation. Narratives for study withdrawal are not described at all, therefore the male-female distribution for study withdrawal could not be determined from the study dossier.
- Denosumab prescription for bone loss associated with aromatase inhibitors used for breast cancer treatment is not approved in Europe. Approval by the FDA and our aim form a large safety analysis set that represent as many denosumab-treated subjects as possible motivated the inclusion of breast cancer subjects in our safety analysis set(36).
- We recognize that the cause of observed sex differences could be a response to the treatment or a difference presentation of the disease. Serious adverse events are not proven to have a causal relationship to the drug. Furthermore, the relatedness to investigational product mentioned in the study reports are still not scientifically proven and therefore not of use.
- There is no collective safety measurement for all system organ classes combined. Because we assess multiple outcomes (each system organ class) in the same safety analysis set, we invite a multiplicity issue. This refers to the potential inflation of finding false positives due to multiple testing.

5.3 Suitability of used safety data

Safety parameter: The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human User (ICH) says all safety parameters are equally important, but do they give the same results when used in sex-specific safety analysis? In this study we deemed serious adverse events as the most suitable for objective sex-specific safety analysis. The incidence rate of adverse events is a valid estimator because it is assumed to be time constant(37). Another type of data that is often presented in scientific articles is time-to-event (or survival) data. This includes information about whether an event occurred, and when it occurred. We found time-to-event data less suitable, because most study reports only included time-to-event data for fatal events. Considering the chronic condition and need for treatment, the short follow up time that was available for analysis (12 months) made time-to-event data less of interest.

System organ class: Modelling the safety profiles on system organ class level might be too broad, because, contrary to the expectation, preferred terms within a SOC are not necessarily medically related (a preferred term is a distinct medical concept that describes a symptom, sight, disease diagnosis, therapeutic indication, investigation, surgical or medical procedure, and medical social or family history characteristic). An additional selection criterium could be to define a threshold for the minimum number of events (or subjects) in a SOC before it is used in analysis. A different approach could be to use another level in the MedDRA (Medical Dictionary for Regulatory Activities) hierarchy instead of the system organ class that is more specific, or to use standardized MedDRA Queries (SMQ's). SMQ is a grouping of terms from one or more SOCs related to a condition of interest.

Clinical study reports as data origin: A disadvantage of using the study reports that are submitted to the medical evaluation board is that it only presents summarized data. Statistical analysis can be performed more easily on data on patient level. However, the power of graphical methods is that they can be used based on summarized data, allowing for intermittent data analysis of (blinded) ongoing trials.

Randomized controlled clinical trials: We used RCTs because of the need of a control group, and because observational studies and real-world post-marketing data often suffers from underreporting. We acknowledge that men included in the clinical may have osteoporosis in a more advanced stadium that the included women. Male osteoporosis is largely underdiagnosed, and male osteoporotic fractures occur about 10 years later in life than women. Due to the advanced age, men may have more comorbidities and suffer more seriously from adverse events, consequently, their mortality is about twice the rate in women(38). Found sex differences in the incidence of serious adverse events could therefore not primarily be due to sex, but due to bias in diagnosis and trial inclusion. This was also seen in a study by Diker-Cohen et al. 2020. Denosumab-treated men were significantly older than treated women and had a lower eGFR and more advanced osteoporosis. These findings suggest that selection bias rather than male gender per se underlies the higher rate of denosumab-associated hypocalcaemia in men(39).

5.4 Comparability of benefits

Efficacy lies outside the scope of our current study and was therefore not assessed. We used the subgroup analysis performed in the GIOP trial (saag et al. 2018), which included both sexes, as a reference. Saag et al. assessed the percent change from baseline in lumbar spine bone mineral density at 12 months. A significant quantitative interaction was observed only in the analysis by sex in the glucocorticoid-initiating subpopulation (vs. the glucocorticoid-continuing subpopulation) (1.2% for men and 3.7% change for women). However, non-significant qualitative interaction testing indicated that there was no evidence that the direction of the denosumab effect differed by gender

in the glucocorticoid-initiating subpopulation. These findings were used to substantiate our assumption that there are no sex differences in denosumab efficacy. There are no mentions of the assessment or comparison of baseline characteristics, pharmacokinetics, dynamics or other clinical endpoints by sex.

Nevertheless, for sound reasoning in the benefit-risk assessment of a drug after sex-specific safety analysis, it is important to rule out any sex differences in the baseline characteristics of a study population, or in the efficacy of the drug, which could be confounders in sex-specific safety analysis. This is where the $2FES_2$ plot would be effective, as it would show potential sex effects in both efficacy and safety.

6. References

1. Vargesson N. Thalidomide-induced teratogenesis: history and mechanisms. Birth Defects Res C Embryo Today. 2015;105(2):140-56.
2. Colombo D, Zagni E, Nica M, Rizzoli S, Ori A, Bellia G. Gender differences in the adverse events' profile registered in seven observational studies of a wide gender-medicine (MetaGeM) project: the MetaGeM safety analysis. Drug Des Devel Ther. 2016;10:2917-27.
3. H. Sundseth PM, K. Semancik. Policy Brief: Sex and Gender in Medicines Regulation2017.
4. (ICH) TICfHoTRfPfHU. CPMP/ICH/363/96 Topic E9 Statistical Principles for Clinical Trials. 2006.
5. (ICH) TICfHoTRfPfHU. E2A Guideline - Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. 1994.
6. (ICH) TICfHoTRfPfHU. Sex-related considerations in the conduct of clinical trials. https://www.ich.org/page/consideration-documents2005.
7. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials. 2012;13:138.
8. (EMA) EMA. European Public Assessment Report Prolia/Denosumab. EPAR. London: Committee for Medicinal Products for Human Use (CHMP); 2018. Contract No.: EMA/CHMP/406583/2018.
9. Aspray TJ, Hill TR. Osteoporosis and the Ageing Skeleton. Subcell Biochem. 2019;91:453-76.
10. Watts NB. Postmenopausal Osteoporosis: A Clinical Review. J Womens Health (Larchmt). 2018;27(9):1093-6.
11. A. Blokstra CAB, H.C. Boshuizen. Vergrijzing en toekomstige ziektelast. Prognose chronische ziektenprevalentie 2005-2025. In: (RIVM) Dniohate, editor. 2007. p. 98.
12. (RIVM) RvVeM. Jaarprevalentie osteoporose. 2019.
13. (RIVM) RvVeM. Famacotherapeutisch kompas: Corticoseroïden, systemisch.
14. Cummings SR. Denosumab for prevention of fractures in postmenopausal women with osteoporosis. The New England Journal of Medicine. 2009:756-65.
15. Briot K, Roux C. Glucocorticoid-induced osteoporosis. RMD Open. 2015;1(1):e000014.
16. Gnant M, Pfeiler G, Dubsky PC, Hubalek M, Greil R, Jakesz R, et al. Adjuvant denosumab in breast cancer (ABCSG-18): a multicentre, randomised, double-blind, placebo-controlled trial. Lancet. 2015;386(9992):433-43.
17. Smith MR, Egerdie B, Hernandez Toriz N, Feldman R, Tammela TL, Saad F, et al. Denosumab in men receiving androgen-deprivation therapy for prostate cancer. N Engl J Med. 2009;361(8):745-55.
18. Agency EM. European Public Assesment Report Prolia/Denosumab. London: Committee for Medicinal Products for Human Use (CHMP); 2010. Contract No.: EMA/21672/2010.
19. Cohen SB. Denosumab Treatment Effects on Structural Damage, Bone Mineral Density, and Bone Turnover in Rheumatoid Arthritis. Arthritis & Rheumatism. 2008:1299-309.
20. Fizazi K, Carducci M, Smith M, Damiao R, Brown J, Karsh L, et al. Denosumab versus zoledronic acid for treatment of bone metastases in men with castration-resistant prostate cancer: a randomised, double-blind study. Lancet. 2011;377(9768):813-22.
21. Henry DH, Costa L, Goldwasser F, Hirsh V, Hungria V, Prausova J, et al. Randomized, double-blind study of denosumab versus zoledronic acid in the treatment of bone metastases in patients with advanced cancer (excluding breast and prostate cancer) or multiple myeloma. J Clin Oncol. 2011;29(9):1125-32.
22. Smith MR, Saad F, Coleman R, Shore N, Fizazi K, Tombal B, et al. Denosumab and bone-metastasis-free survival in men with castration-resistant prostate cancer: results of a phase 3, randomised, placebo-controlled trial. Lancet. 2012;379(9810):39-46.
23. Saag KG, Pannacciulli N, Geusens P, Adachi JD, Messina OD, Morales-Torres J, et al. Denosumab Versus Risedronate in Glucocorticoid-Induced Osteoporosis: Final Results of a Twenty-Four-Month Randomized, Double-Blind, Double-Dummy Trial. Arthritis Rheumatol. 2019;71(7):1174-84.

24. Phillips R, Sauzet O, Cornelius V. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. BMC Med Res Methodol. 2020;20(1):288.
25. Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
26. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
27. Cooper AJ, Lettis S, Chapman CL, Evans SJ, Waller PC, Shakir S, et al. Developing tools for the safety specification in risk management plans: lessons learned from a pilot project. Pharmacoepidemiol Drug Saf. 2008;17(5):445-54.
28. Michel C, Scosyrev E, Petrin M, Schmouder R. Can Disproportionality Analysis of Post-marketing Case Reports be Used for Comparison of Drug Safety Profiles? Clin Drug Investig. 2017;37(5):415-22.
29. Alvarez-Requejo A, Carvajal A, Begaud B, Moride Y, Vega T, Arias LH. Under-reporting of adverse drug reactions. Estimate based on a spontaneous reporting scheme and a sentinel system. Eur J Clin Pharmacol. 1998;54(6):483-8.
30. Harrer M, Cuijpers, P., Furukawa, T.A., & Ebert, D.D. Publication bias. Doing Meta-Analysis with R: A Hands-On Guide

London: Chapmann & Hall/CRC Press.; 2021.
31. Orwoll E, Teglbjaerg CS, Langdahl BL, Chapurlat R, Czerwinski E, Kendler DL, et al. A randomized, placebo-controlled study of the effects of denosumab for the treatment of men with low bone mineral density. J Clin Endocrinol Metab. 2012;97(9):3161-9.
32. Saag KG, Wagman RB, Geusens P, Adachi JD, Messina OD, Emkey R, et al. Denosumab versus risedronate in glucocorticoid-induced osteoporosis: a multicentre, randomised, double-blind, active-controlled, double-dummy, non-inferiority study. Lancet Diabetes Endocrinol. 2018;6(6):445-54.
33. Dekker M, de Vries ST, Versantvoort CHM, Drost-van Velze EGE, Bhatt M, van Meer PJK, et al. Sex Proportionality in Pre-clinical and Clinical Trials: An Evaluation of 22 Marketing Authorization Application Dossiers Submitted to the European Medicines Agency. Front Med (Lausanne). 2021;8:643028.
34. De Martinis M, Sirufo MM, Polsinelli M, Placidi G, Di Silvestre D, Ginaldi L. Gender Differences in Osteoporosis: A Single-Center Observational Study. World J Mens Health. 2021;39(4):750-9.
35. Foundation) IIO. Epidemiology, Burden and Treatment of Osteoporosis for 29 European countries (2021) 2021 [Available from: https://www.osteoporosis.foundation/facts-statistics/key-statistic-for-europe.
36. Medicines Guide: Prolia (denosumab) Full Prescribing Information, (2020).
37. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. Pharm Stat. 2016;15(4):297-305.
38. Vescini F, Chiodini I, Falchetti A, Palermo A, Salcuni AS, Bonadonna S, et al. Management of Osteoporosis in Men: A Narrative Review. Int J Mol Sci. 2021;22(24).
39. Diker-Cohen T, Amitai, O., Shochat, T., Shimon, I., & Tsvetov, G. Denosumab-associated hypocalcemia: Does gender play a role? Maturitas. 2020;142:17-23.
40. Amgen. A Study to Evaluate Denosumab in the Treatment of Postmenopausal Osteoporosis: Study results www.clinicaltrials.gov2004 [updated July 7, 2020. Available from: https://clinicaltrials.gov/ct2/show/results/NCT00089791?term=freedom&cond=Osteoporosis%2C+Postmenopausal&draw=2&rank=1&view=results#wrapper.
41. Agency) OFEM. Draft presentation: Summary of product characteristics. Medical information, compliance and consistency.
42. Agency) SAGEM. SmPC training presentation: Section 4.8: Undesirable effects.
43. Agency EM. Opening up clinical data on new medicines 2016 [Available from: https://www.ema.europa.eu/en/news/opening-clinical-data-new-medicines.
44. Clinical data [Internet]. Available from: https://clinicaldata.ema.europa.eu/web/cdp/home.

45. European database of suspected adverse drug reaction reports [Internet]. Available from: https://www.adrreports.eu/en/eudravigilance.html.
46. Medication Guides [Internet]. Available from: https://www.adrreports.eu/en/eudravigilance.html.
47. FDA Adverse Event Reporting System (FAERS) Public Dashboard [Internet]. Available from: https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard.
48. Langdahl BL, Teglbjaerg CS, Ho PR, Chapurlat R, Czerwinski E, Kendler DL, et al. A 24-month study evaluating the efficacy and safety of denosumab for the treatment of men with low bone mineral density: results from the ADAMO trial. J Clin Endocrinol Metab. 2015;100(4):1335-42.
49. Sugimoto T, Matsumoto T, Hosoi T, Miki T, Gorai I, Yoshikawa H, et al. Three-year denosumab treatment in postmenopausal Japanese women and men with osteoporosis: results from a 1-year open-label extension of the Denosumab Fracture Intervention Randomized Placebo Controlled Trial (DIRECT). Osteoporos Int. 2015;26(2):765-74.
50. Cohen SB, Dore RK, Lane NE, Ory PA, Peterfy CG, Sharp JT, et al. Denosumab treatment effects on structural damage, bone mineral density, and bone turnover in rheumatoid arthritis: a twelve-month, multicenter, randomized, double-blind, placebo-controlled, phase II clinical trial. Arthritis Rheum. 2008;58(5):1299-309.
51. Ellis GK, Bone HG, Chlebowski R, Paul D, Spadafora S, Smith J, et al. Randomized trial of denosumab in patients receiving adjuvant aromatase inhibitors for nonmetastatic breast cancer. J Clin Oncol. 2008;26(30):4875-82.
52. Kendler DL, Roux C, Benhamou CL, Brown JP, Lillestol M, Siddhanti S, et al. Effects of denosumab on bone mineral density and bone turnover in postmenopausal women transitioning from alendronate therapy. J Bone Miner Res. 2010;25(1):72-81.
53. Tsvetov G, Amitai O, Shochat T, Shimon I, Akirov A, Diker-Cohen T. Denosumab-induced hypocalcemia in patients with osteoporosis: can you know who will get low? Osteoporos Int. 2020;31(4):655-65.
54. (ICH) TICfHoTRfPfHU. Efficacy - M4E(R2) Guideline. Revision of M4E Guideline on Enhancing the Format and Structure of Benefit-Risk Information in ICH. 2016.
55. (ICH) TICfHoTRfPfHU. E3 Guideline - Structure and Content of Clinical Study Reports. 1995.
56. Chuang-Stein C, Xia HA. The practice of pre-marketing safety assessment in drug development. J Biopharm Stat. 2013;23(1):3-25.
57. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods. 2016;7(1):55-79.
58. Harrer M, Cuijpers, P., Furukawa, T.A., & Ebert, D.D. Between-study heterogeneity. Doing Meta-Analysis with R: A Hands-On Guide. London: Chapmann & Hall/CRC Press.; 2021.
59. Harrer M, Cuijpers, P., Furukawa, T.A., & Ebert, D.D. Subgroup Analyses. Doing Meta-Analysis with R: A Hands-On Guide. London: Chapmann & Hall/CRC Press.; 2021.

Appendix A identified data sources

Here we discuss the usability of each identified data source, by determining what kind of safety data it provides, how aggregated the presented data is (patient level data or more summarized into counts or filtered by a minimum incidence in either treatment arm) and if the data is presented unfiltered (e.g., only presenting pre-specified safety outcomes).

A1 Clinical trial registries

Clinical trials with denosumab as the investigational product are registered in freely accessible online databases www.clinicaltrials.gov, the EU Clinical Trials Register and the International Clinical trials registry platform (ICTRP).

These platforms offer search tools to filter out phase 3 and 4 trials with denosumab as the investigational product. Details of the study such as inclusion and exclusion criteria, study design, arms and interventions are described. Summaries of the participant flow, baseline characteristics and primary and secondary outcome measures of the trials are presented. Hyperlinks to related publications of the results are available. This makes the clinical trial registries very helpful in the selection of trials with safety data from a representative study population.

Clinicaltrials.gov presents adverse event data at the preferred term level, in two column tables, one column for the control group and one column for the group treated with the investigational product. All-cause mortality of presented as an affected / at risk (%) ratio a.k.a. subject incidence. However, often this section is not filled out. Subject incidences of (serious) adverse events are presented by preferred term only. A difficulty is that there is no mention of subject incidence by system organ class. The cumulative subject incidence at the end of the trial is reported, not per year, while the lengths or the trials can vary between 1 to 10 years. This makes pooling of the data from multiple trials or comparisons between trials less accurate. Additionally for some trials a frequency threshold for reporting events is used, making these reported events incomplete and insufficient for comparison of the complete safety profile (e.g. Other (Not Including Serious) Adverse Events in the FREEDOM trial)(40). Most importantly, when both sexes are included in a study, the clinical trial registries only report safety data per treatment arm, not per sex, making these data unfit for sex-specific safety analysis.

Disclaimer: Not all listed studies are regulated and/or reviewed by the U.S. FDA or other governmental entities. The National Library of Medicine (provider of CT.gov) does not verify the scientific validity or relevance of the submitted information beyond a limited quality control review for apparent errors, deficiencies or inconsistencies.

A2 Regulatory authorities

A2-1 Summary of Product Characteristics

The Summary of Product Characteristics (SmPC) is a legal document formatted by the EMA that is approved as part of the marketing authorisation of each medicine. It includes information about the use of the medicine, the benefits and risks of the medicine, individualised care and pharmaceutical information. The document is openly accessible, but mostly meant for healthcare professionals to prescribe the medicines safely and effectively(41).

Children and elderly are a specific subpopulation that are considered in need of specificity in the use of the medicine. It is mandatory to address appropriate use, specific risks or clinically relevant differences in these populations in the SmPC. Men or women are not mentioned as a possible subpopulation.

Section 4.8 “Undesirable effects’ contains a summary of the safety profile that mentions the most frequently seen adverse reactions(42). It should include all adverse reactions from clinical trials, post-authorisation safety studies and spontaneous reporting. These pooled data are presented in a table by system organ class with a frequency category of very common to very rare (Prolia SmPC 2018 p.6(8)). Selected adverse reactions are highlighted and the trials and frequency in which they occur are mentioned. made about the possible effect of sex on the observed adverse event frequencies nor is there a sex-specific analysis. Efficacy and safety data per trial are mentioned, however upon closer inspection, only efficacy and the effect on reduction of fracture incidence are discussed.

All phase 3 and 4 studies in the Prolia SmPC were also found in the search through clinical trial registries. Overall, the information found in the SmPC gives information on which trials were deemed pivotal in the regulatory decision-making process and why, but the SmPC does not provide and is neither fit for sex-specific safety analysis.

A2-2 European Public Assessment Report

The European Public Assessment Report (EPAR) is published for every medicine and can be seen as a more extensive version of the SmPC. It contains more in-depth information about the studies referenced for market authorisation, about the periodic safety updates, and about how this was assessed by the EMA.

The EPAR mentions safety concerns categorised as important identified risks, important risks and missing information, which could be used as pre-determined safety signals which will need to be assessed more in depth when performing sex-specific safety analysis (table 4). Interestingly, risk differences in males and females, is not mentioned under missing information.

Again, only summarized safety information is presented, without sex-specific data or analysis. Therefore, the EPAR also cannot aid in the comparison of denosumab safety profiles between the sexes.

Table 4. Safety concerns mentioned in the European Public Assessment report of 2018 (8).

Important identified risks	hypocalcaemia, skin infection leading to hospitalization, osteonecrosis of the jaw, hypersensitivity reactions, atypical femoral fracture, musculoskeletal pain
Important potential risks	fracture healing complications, infection, cataracts in men with prostate cancer receiving androgen deprivation therapy, cardiovascular events, malignancy, immunogenicity, osteonecrosis outside the jaw including external auditory canal, hypercalcemia following treatment discontinuation in patients with growing skeletons
Missing information	risks with pregnancy/lactation, use in paediatric patients, use in patients with hepatic impairment, potential adult off-label use

A2-3 EMA transparency initiative

Since October 2016 the EMA states to give open access to clinical reports for new medicines for human use(43). Clinical reports contained in all marketing authorisation applications submitted to the EMA are available through a publicly accessible EMA database(44). This database is gradually growing with clinical trial data. Despite Prolia being authorised by the EMA since 2016, at this moment no documents containing information on Prolia (the product name), denosumab (the active substance) or M05BX04 (the anatomical therapeutical chemical (ATC) code) are found in the database.

The transparency initiative is a progressive approach that could facilitate re-analysis and meta-analysis of clinical trial data. It helps health care professionals, researchers and the public to investigate trial results and understand the underlying data that led to regulatory decisions. However, this database cannot facilitate or contribute to our objective to perform sex-specific safety analysis on denosumab yet.

The regulatory pharmacovigilance system of the European Union, named EUdravigilance, designed for the reporting of adverse events. The website states adverse events in both pre- and post-authorisation phases are collected by EUdravigilance. The Clinical Trial Module collects data on suspected unexpected serious adverse reactions (SUSARs) reported by the sponsors of interventional clinical trials. Post-authorisation data that is collected is broader in nature, because both serious and non-serious AE data is collected from healthcare professionals and patients spontaneous reporting, from non-interventional studies and from worldwide scientific literature. An online browse tool allows searching for web reports on suspected side effects for a medicine or active substance(45). They all contain: the number of individual cases, presented by age group, sex and geographic origin; the number of cases by reaction group (e.g., SOC) presented by age group, sex, seriousness, reporter group and geographic origin; and more. Unfortunately, the cases in the web reports only show adverse events related to spontaneous cases since the medicine was authorised for use in the European Economic region. Therefore, clinical trial data will not be included. Despite reports being presented by sex, spontaneous reporting is highly subjected to reporting bias and the data does not consider the number of medicine users. Also, this database is young and reports only go as far back as February 2021. Therefore, these data cannot be used to assess sex-specific medicine safety.

A2-4 FDA

The U.S Food and Drug Administration is the American version of the EMA. On the FDA website all medication guides can be searched(46). The Medication guide is comparable to the SmPC and EPAR, including prescription indications and contraindications, efficacy and safety information. The most common (serious) adverse events are listed in total and separately for patients with PMO or GIOP, for male users, and for patients with bone loss receiving androgen deprivation therapy or aromatase inhibitor therapy for prostate and breast cancer(36). The guide included risk summaries in specific populations, e.g., section 8.1 ‘Females and males of reproductive potential’, but it is only stated that Prolia can cause foetal harm, which is the same information stated in section 8.1 ‘Pregnancy’. It cannot be concluded from this document if the FDA did not find sex-specific safety issues, or that sex-specific analysis was not performed. In section 12.3 ‘Pharmacokinetics’ it is mentioned that the mean serum denosumab concentration-time profiles were not affected by gender, race and age.

The FDA Adverse Event Reporting System can be compared to EUdravigilance, monitoring the safety of a medicine through adverse event reports they receive. FAERS data only includes post-marketing surveillance data. This causes some limitations, as stated by the FDA itself: potential submission of incomplete, inaccurate, unverified information; the incidence or prevalence of an event cannot be determined from this reporting system alone due to potential under-reporting of events and lack of information about the frequency of use. The case count for Prolia users goes as far back as 2010. Data as of December 31, 2021, shows a great difference of almost 100,000 reports by females vs. 8,059 by males (vs. 10,649 not specified). This implies either that women have more side effect, that women report spontaneously more often, or that the number of users or time of use is bigger/longer for women(47).

It can be concluded that the FDA has similar available information as the EMA and therefore the FDA database cannot facilitate or contribute to our objective to perform sex-specific safety analysis on denosumab either.

A3 Published literature

Published literature about denosumab studies were searched via a selection of bibliographical databases PubMed, Cochrane Central Register of Controlled Trials (CENTRAL) and Google Scholar. As mentioned in section 3.1.1. 'Clinical trial registries', many different types of studies have been performed and published, single arm and multi-arm trials, observational and intervention studies, dose finding studies etc. The trial registries provided hyperlinks to published articles related to the clinical trial. There are not many trials, or reviews, which discuss sex-specific denosumab safety. Most trials are single sex, and the majority is female. The scarce trials on osteoporotic men treated with denosumab emphasises the little amount of data about male osteoporosis and denosumab effects of increasing bone mineral density that is comparable to effects seen in postmenopausal women (17, 48). A review from Vescini et al. 2021 mentions the underdiagnosis of male osteoporosis and a higher male fracture-related mortality rate(38). Trials that include both sexes stratify the control vs. treatment arm, but they do not present or mention sex-specific data or analysis(32, 49).

Most published trial articles present the frequency of adverse events, serious adverse events, events leading to discontinuations, deaths and selected events of interest in a table, including the placebo or active comparator group and the denosumab group if present (50, 51). Sometimes a third column presents the p-value based on Fisher's exact test(52). However, often the presented contingency table does not display all system organ classes, only preferred terms that were observed often, or only pre-specified safety outcomes. Also, the way of reporting safety across papers is not uniform, which impedes efficient safety analysis that combines safety data from multiple trials (exposure adjusted incidence, subject incidence, event incidence).

The Dutch government as well as the International Osteoporosis Foundation provide yearly osteoporosis prevalence data per sex, per country and for a total of 29 European countries, registered in healthcare databases(12, 35). The number of deaths is also mentioned, however not per sex. These sources do not provide information about which osteoporotic treatment is used nor does it mention safety (signals).

To summarize, published literature shows that osteoporosis occurs mostly in women, touches upon men that are underdiagnosed and that this could be the underlying reason for observed sex-differences in retrospective safety analysis of real-life data from a health maintenance organisation(53). However, these publications do not display enough unfiltered, in-depth data to be of use in sex -specific safety analysis.

A4 Regulatory submissions

Pharmaceutical companies that want to receive European or national marketing authorisation submit a dossier, also known as the Common Technical Document (CTD) to be assessed by regulatory authorities. the European Medicines Agency grants European authorisation, while e.g., the Dutch Medicines Evaluation Board grants national authorisation. A dossier must meet certain requirements concerning content and layout. It must include 5 modules (fig. 7):

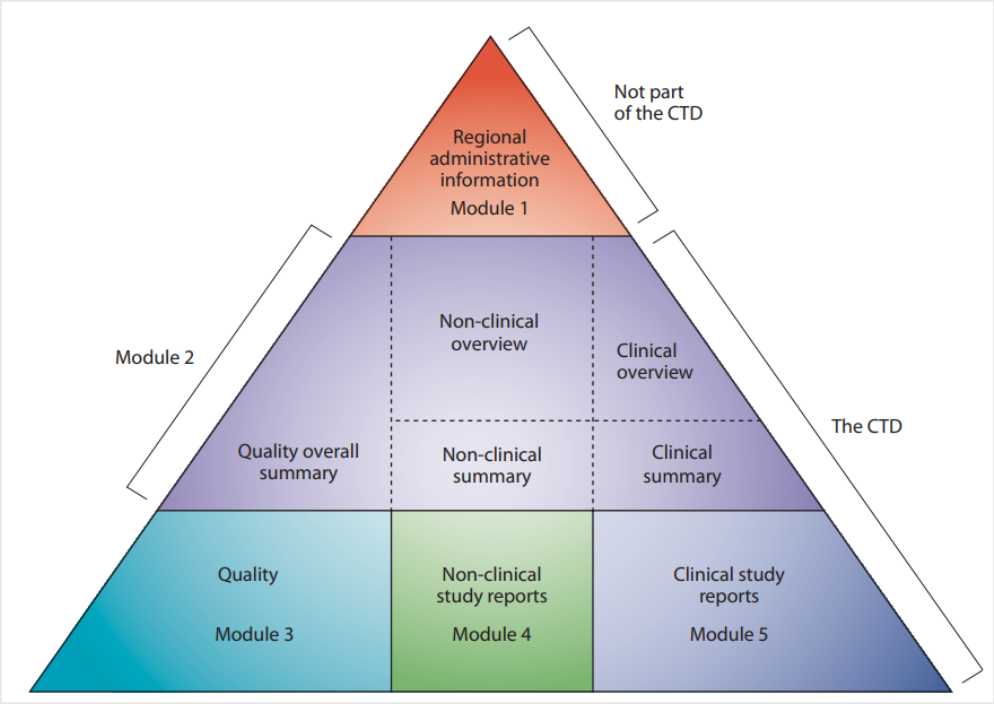

Figure 7. The common technical document (CTD) triangle. THE CTD is organized into five modules. Module 1 is region specific and modules 2, 3, 4 and 5 are intended to be common for all regions (figure used from the ICH website).

- Module 1 contains administrative data, including the SmPC. This document holds the most important scientific data about the medical products for doctors and pharmacists.
- Module 2 contains the summaries of chemical-pharmaceutical, pharmacological-toxicological and the clinical-pharmacological dossier
- Module 3 all chemical-pharmaceutical data: about the composition and preparation and quality control of the medicinal product
- Module 4 contains all pharmacological toxicological data: about animal experiments relating to toxicity and the mechanism of action of the medicinal product
- Module 5 contains all clinical-pharmacological data: about the efficacy and safety of the medicinal product in humans

A4-1 Clinical overview

The clinical overview (CTD section 2.5) is a summary in module 2 of the CTD that provides assessment of the clinical data. It presents the strengths and limitations of study results analyses the benefits and risks of a medicine and describes how the study results support decisions made in the prescribing information. Clinical overviews are not publicly accessible. The database of the Dutch regulatory authority was used to access the and is not publicly available CTD section 2.5.5 zooms in on safety data and according to the ICH guideline should provide critical analysis of adverse effects, relevant animal toxicology and product quality information, the nature of the patient population and extent of treatment exposure, common-, and non-serious and serious adverse events, similarities and differences in results among studies, any differences in rates of AEs in population subgroups and more(54). Gender or sex is not mentioned as a subgroup. As this is an overview, emphasis is placed on cross-referencing to other modules for tabulated data on AE rates.

The overview of clinical pharmacology states that the pathophysiology of osteoporosis and the pharmacokinetics and pharmacodynamics of denosumab is not significantly affected by sex, weight or disease state, but does not provide the data used to reach this conclusion. Therefore, the targeted

dosing regime is the same for subjects with PMO and bone loss associated with hormone ablation therapy breast or prostate cancer. The overview of Safety does not mention any sex-specific analysis or decision making. The clinical overview is difficult to use, because there is not just one overview that gets updated when new data becomes available. Instead, as new studies are conducted, sometimes a new updated overview is made and added to and stored with the CTD of that study. As all studies are not stored together, it is hard to know or if there are newer studies conducted with an updated clinical overview. The most recent one found dates to May 2014. Overall, the clinical overview does not provide the safety data needed for the safety analysis we want to perform.

A4-2 Clinical summaries

The clinical summary (CTD section 2.7) is another summary in module 2 that focusses on data summarization and integration.

The ICH guideline states: It is usually useful to examine more closely the more common adverse events that seem to be drug related (e.g., those that show that a dose response and/or a clear difference between drug and placebo rates) for relationship to relevant factors, among others: age, sex and race. Rigorous statistical evaluation of the possible relationship of specific adverse events to these factors are often unnecessary. When there is no evidence of a significant relationship upon display and inspection of the data, no further analysis of these factors is necessary. Tabulated presentation of events of interest per treatment arm are presented only for the data retrieved from the latest trial, in this case Orwoll2012. Thorough between-study analysis is performed as well as safety in special groups and situations, but mostly without the data that lead to the conclusions. There is no mention of between-sexes analysis.

A4-3 Clinical study reports

Module 5 contains the clinical study report, which adheres to a clear structure and content defined by the ICH. The E3 guideline states that the clinical study report should be an integrated, full report of an individual study of any drug or treatment conducted in patients. It includes the clinical and statistical description, presentations and analyses, incorporating tables and figures and appendices containing the protocol, sample case report forms, investigator related information, information related to the investigational product, including active control/comparators, patient data listings and technical statistical documentation and details such as derivations, computations, analyses and computer output(55). These reports are not publicly available. The clinical study reports of the various trials selected through the clinical trial registries and SmPC/EPAR are retrieved from the Dutch regulatory authority's database.

In contrast to the clinical summary and overview which contains combined information on multiple trials, the study report contains all information on a single trial. Safety data from the clinical study report includes various tables presenting the incidence of (serious) adverse events at preferred term and system organ class level, discontinuations, withdrawals, deaths and narratives. Narratives provide details on time-to-event since the first dose of investigational product, the outcome of the event (e.g., fatal, resolved, drug withdrawn etc.) and subject information such as age, sex and comorbidities. Most trials occur in a single sex making it impossible to perform sex-specific analysis. Often the reports refer to other trials to define expected risks or justify chosen cut-off margins (Clinical study report Saag2018 p.49/50). Results are however not rarely compared to or combined with data from previous trials. This could be because the lack of access to proper data necessary for pooled analysis. Saag et al. 2018 does include both sexes and presents data per sex, with further statistical analysis on efficacy data but not on safety data. Brief mentioning of comparisons across trials are made: *"Denosumab exposure in subjects with GIOP was consistent with that observed*

previously in subjects with postmenopausal osteoporosis (Study 20030216)" Clinical study report Saag2018, p.85.

While the clinical study report often does not mention sex-specific safety, because most trials are only conducted in a single sex, the data is specific enough to be used for analysis. By combining data from multiple reports of randomized controlled denosumab trials, we expect to be able to analyse sex-specific safety.

A4-4 Integrated Summary of Safety

CTD section 5-3-5-3 is the Integrated Summary of Safety and contains reports of analyses of data from more than one study. The integrated summary of safety of the GIOP trial reports summaries of the subject incidence of treatment-emergent (serious) adverse events from each of the five denosumab studies used in this research: the GIOP trial, the male osteoporosis (ADAMO) trial, the postmenopausal osteoporosis (FREEDOM) trial, the HALT trial (men with prostate cancer) trial and the AIT trial (aromatase inhibitor therapy) (women with breast cancer), side by side. There are various tables which summarize differently: per trial, per clinical condition and per sex. Not only the safety parameters from each trial are presented side by side, also the subject disposition, baseline demographics, other adverse events of interest, clinical laboratory evaluations of calcium and vital signs are presented. It does not provide any further analysis. The integrated summary of safety presents an easy overview of all safety data that we want to use in our sex-specific safety analysis and is therefore a valuable data source. As the combined baseline characteristics are also presented by sex, this information could be analysed for sex differences as well and taken into consideration in the sex-specific safety analysis.

Appendix B data visualization methods

Table 5. Ranking of the visualization methods based on ease of implementation, ease of comprehension and aesthetics. Appendix B presents imaging and discussion of each method separately. SOCs = System Organ Classes, names are abbreviated. Full names of the system organ classes can be found in Appendix C1.

Rank	Method	Depicted safety information	Main advantage	Main disadvantage	Detected sex-specific safety signals: SOCs (Subject to viewer)
1	Volcano plot	Relative risk + Fisher's exact p-value	Detects SOCs based on p-value	No direct comparison between sexes	Injury
2	Forest plot	relative risk per sex + CI + (weighted) average relative	Detects SOCs based on CI	Overly large plot needed to display all SOCs Statistical analysis	Injury
3	Dot plot	Absolute subject incidence + relative risk + CI	Detects SOCs based on CI	Interpretation requires two plots	Gastrointestinal Injury Respiratory
4	Contingency table	Absolute subject incidence numbers and percentage) per	Straightforward presentation of all SAE counts	Only presents numbers Only presentation, no	Injury Cardiac Gastrointestinal Respiratory
5	Heatmap	Relative SOC incidence (%) per sex per treatment arm	Easy to spot major outliers	Relative incidence is misleading for interpretation Only	Respiratory Renal Injury Gastrointestinal
6	Stacked bar chart	Relative SOC incidence (%) per sex per treatment arm	Easy to spot major outliers	No common base Presentation of relative incidence is misleading	Gastrointestinal Injury Vascular respiratory

B1 Volcano plot

A volcano plot is a type of scatter or dot plot that shows the statistical significance (p-value) on the y-axis against the magnitude of change measured between two groups (risk ratio). The signature expected shape of the volcano plot is a parabola, when the data is normally distributed, due to the -log10 transformation of the y-axis' scale. Greater risk differences are expected to have higher statistical significance. Points to the right of the dashed line at x-intercept 1 mean that there is a relative higher incidence in the male group, points to the left a higher incidence in the female group (fig. 8). P-values are derived from a two-tailed Fisher's exact test conducted for each system organ class, one in the control arm and one in the denosumab arm. The significance threshold of 5% is used for safety signal detection. SOC's plotted above the threshold may indicate a safety signal: information on a new or known adverse event that may be caused by a medicine and requires further investigation.

Implementation of the volcano plot requires some extra analysis to provide the statistical analysis data, but the aesthetics are pleasing and clear and easy to comprehend. The volcano plot focusses on the screening of the adverse event profile per sex and on safety signal detection(56). It visualizes more than the subject incidence, which is valuable in the comparison between the sexes. We conclude from this visualization method that a safety signal arises from the system organ class Injury, in the female subset of the study population. This implicates a difference with the male subset and therefore needs further investigation into a potential sex- effect on serious adverse event incidence in this SOC.

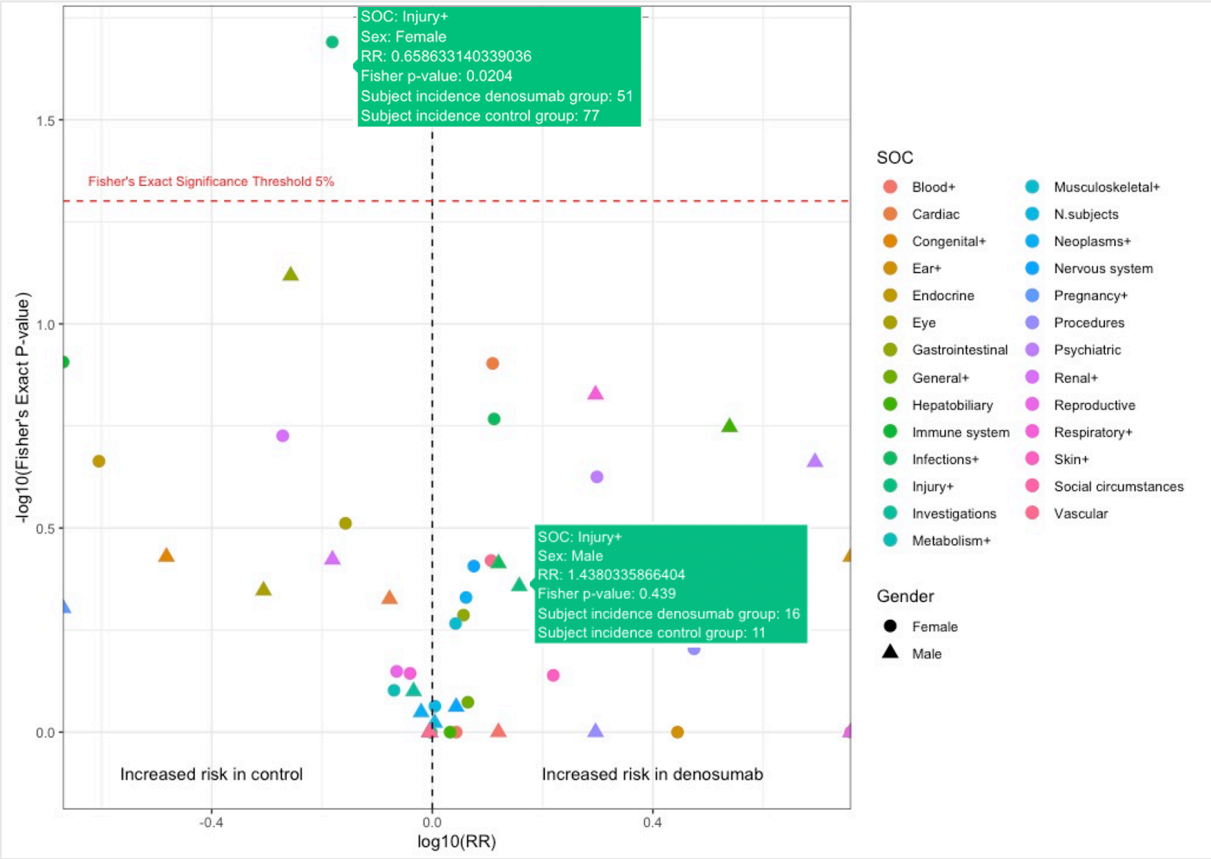

Figure 8. Volcano plot displaying the risk ratio of control vs. denosumab treatment for men and women separately, combined with statistical significance of the risk ratio determined by Fisher's exact test. SOC = System Organ Class, RR = relative risk. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

Advantages
Easy to spot safety signals objectively with the significance threshold
Plotting based on statistical significance of risk ratios
Safety signal detection threshold can be altered to fit the research question (e.g., Bonferroni correction)
It is both a data visualization method as well as a quantitative statistical method
Disadvantages
No direct comparison between the sexes and the treatment arms simultaneously
Generating the statistical analysis data used for plotting is laborious (subjective to the proficiency of the researcher)
The number of SOCs makes differentiation between each colour (which represents a system organ class) challenging

B2 Forest plot

The forest plot is the most common way to visualize meta-analysis results. Meta-analysis combines, summarizes and interprets available evidence in a quantitative way, across multiple studies.

The forest plot uses dots to present the relative risk of a serious adverse event under denosumab treatment vs. control treatment across the x-axis. The lines running though the points present the 95% confidence interval of the relative risk point estimates. The pooled overall relative risk is presented as a diamond shape at the bottom of the plot. The centre of the diamond represents the pooled point estimate, and the horizontal tips represent the confidence interval. The width of the diamond reflects the confidence interval. The point estimates of the relative risks are surrounded by a square. The size of the square represents the weight (fig. 9). When pooling the relative risks in meta-analysis, studies with a higher precision, meaning a smaller standard error, are assigned a greater weight. The bigger the square, the greater the weight, which is also presented in the outer right column. The forest plot offers the possibility to present additional information in columns next to the forest plot itself. Here we chose to present the system organ class, sex, serious adverse event incidence and the group total (per SOC per treatment arm), the relative risk, the 95% confidence interval and the weight of each SOC.

All relative risk point estimates have confidence bands that that cover the value 1, meaning that the risk of an SAE in control and denosumab treatment is similar. Only the SOC injury in the female subset of the study population it does not include 1, this presents a sex-specific safety signal that arises from that system organ class.

According to the random effects model, the amount of variation in the data is low and caused by sampling errors, and not by between-study heterogeneity (see chapter 3.3.2.2: meta-analysis).

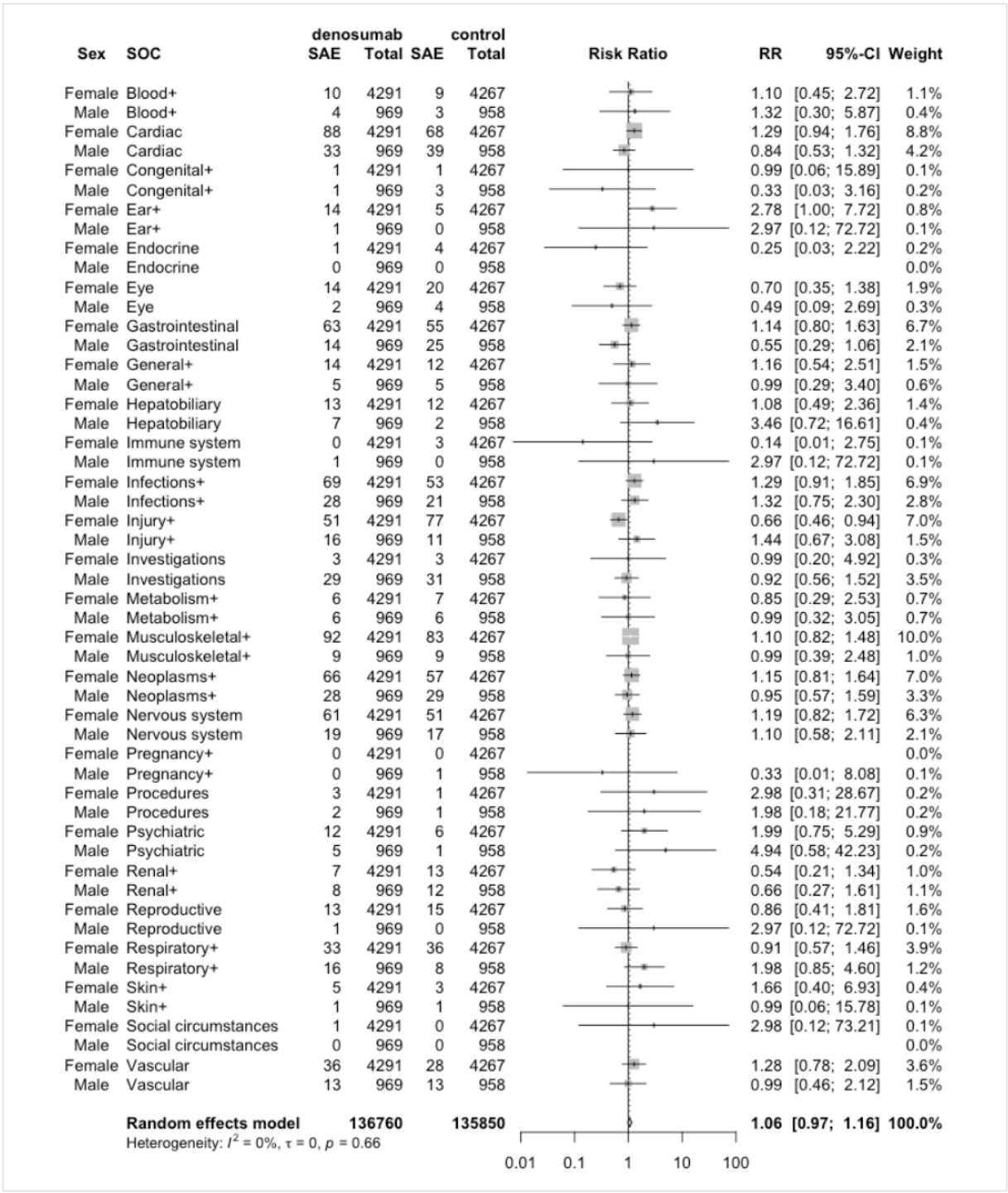

Figure 9. Forest plot of relative SAE risk of denosumab vs. control treatment with pooled SAE data of 5 studies, per system organ class, per sex. SOC = System Organ Class, RR = relative risk or risk ratio. CI = confidence interval. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

Advantages
Many descriptive statistics can be added to the plot
Easy implementation
The forest plot presents a relative risk per SOC per sex, but also an over-all (pooled) relative risk of all the SOCs combined
It is both a data visualization method as well as a quantitative statistical method
Disadvantages
Relative risks of control vs. denosumab can only be analysed per sex. Therefore, forest plot does not directly assess potential sex differences.
The pooling of the data for the random effects model that calculates an over-all effect size (relative risk) cannot be used. It adds up the group total of each SOC, while this is the same group of subjects.
The plot falls short on a dimension, because we want to compare both control vs. treatment and men vs. women, preferably at the same time.
The ease of comprehension is moderate. Safety signals for treatment effects can be spotted easily, if the researcher knows that a confidence band that does not include 1 is seen as a safety signal.
The generated plot is very large in order to present all SOCs, which makes it look cluttered.

B3 Dot plot

The dot plot displays the absolute abundance of subjects with an SAE in one of the four groups (males and females in the denosumab and treatment arm). On the y-axis, each SOC is presented, while the dots present the percentages of subjects with a serious adverse event, measured on the x-axis. Colours correspond to the sexes, while the shape of the points correspond with the treatment (fig. 10). Its focus lies on screening and signal detection(56). The objective is to compare all four groups, but to keep the display clear, the plot is split up in the female and male subset of the study population.

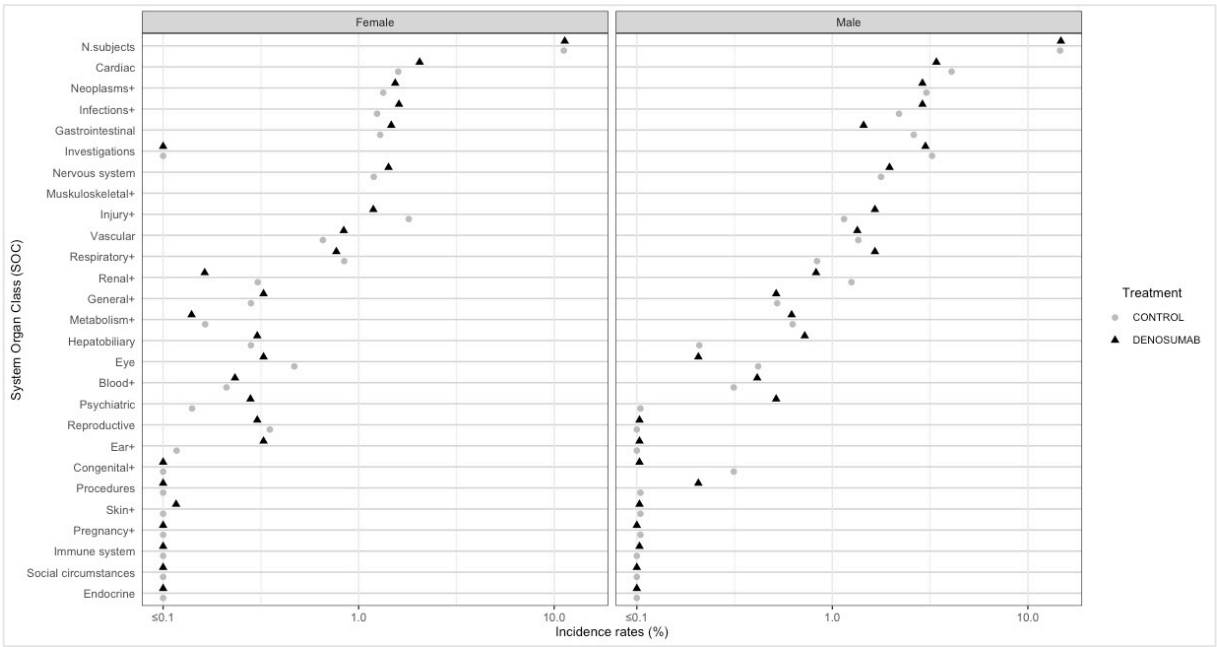

Figure 10. Dot plot of percentage of subjects with serious adverse events per system organ class in the control and denosumab treatment arm, for men and women separately. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

After displaying the absolute abundances, the relative risks have been calculated and plotted in a relative risk dot plot (fig. 11). Again, these control vs. denosumab relative risks of a serious adverse event are plotted into two separate plots to keep the presentation clear.
When the 95% confidence bands of a relative risk do not overlap with the vertical dotted line at value 1 on the x-axis, it implies that there is a significant risk difference (like the forest plot), which is detected as a safety signal. A relative risk above 1 means a higher incidence in de denosumab treatment arm, below 1 means a higher incidence in the control treatment arm. Based on the first dot plot, safety signals are subjectively detected from three system organ classes: gastrointestinal, injury, respiratory. When using the second dot plot, only the system organ class injury would present a safety signal.

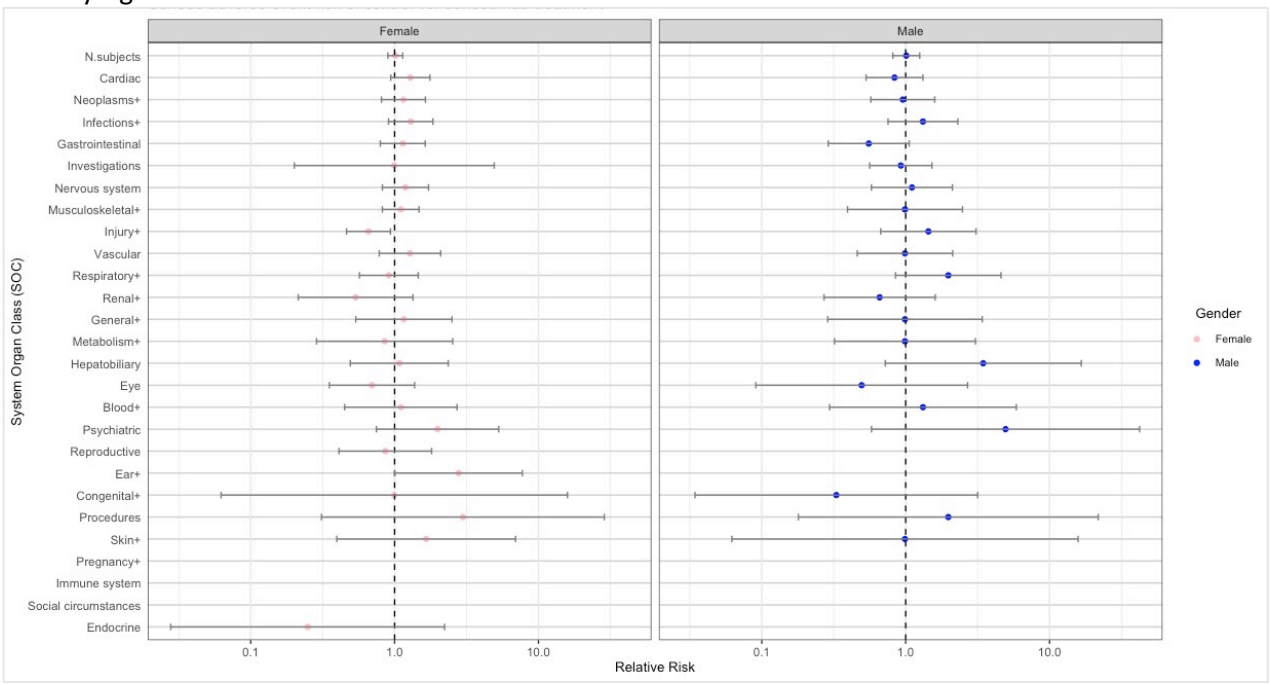

Figure 11. Dot plot of the relative risk of a serious adverse event subject incidence in control vs. denosumab treatment, for men and women separately. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

Advantages
Implementation is easy once the researcher knows how to code and make the plot
Clear overview of each SOC
Incidence pattern differences between sexes can be spotted
It is both a data visualization method as well as a quantitative statistical method, because a relative risk is calculated and visualized in a second plot
Disadvantages
Objective detection of (sex-specific) safety signals is challenging because two separate plots are made
The plots lack a dimension, because we want to compare both control vs. treatment and men vs. women simultaneously
The ease of comprehension is moderate. Safety signals for treatment effects can be spotted easily, if the researcher knows that a confidence band that does not include 1 is seen as a safety signal.

B4 Heatmap

The heatmap displays the relative abundance of SAEs. The treatment arm and sex are on the x-axis and the system organ class is displayed on the y-axis. The values that are plotted are displayed as tiles. The intensity of the colour of each tile corresponds with the relative abundance value. A darker colour means a higher relative abundance of SAEs of that SOC in that treatment group (fig. 12). The

implementation of the heat map is easy, as it uses the information from the contingency table and converts it into a colour gradient. The shift in colour intensity between the control and denosumab treatment group can be used to compare the male and female subset of the study population. To use this method to detect sex-specific safety signals is very subjective. Also, because the presentation could be misleading in the comprehension of the plot, because rectangles with the same colour intensity do not imply that the number of subjects counted is the same. Through this subjective way of sex-specific safety analysis, six safety signals were detected in the system organ classes: respiratory, renal, injury, gastrointestinal, ear and cardiac.

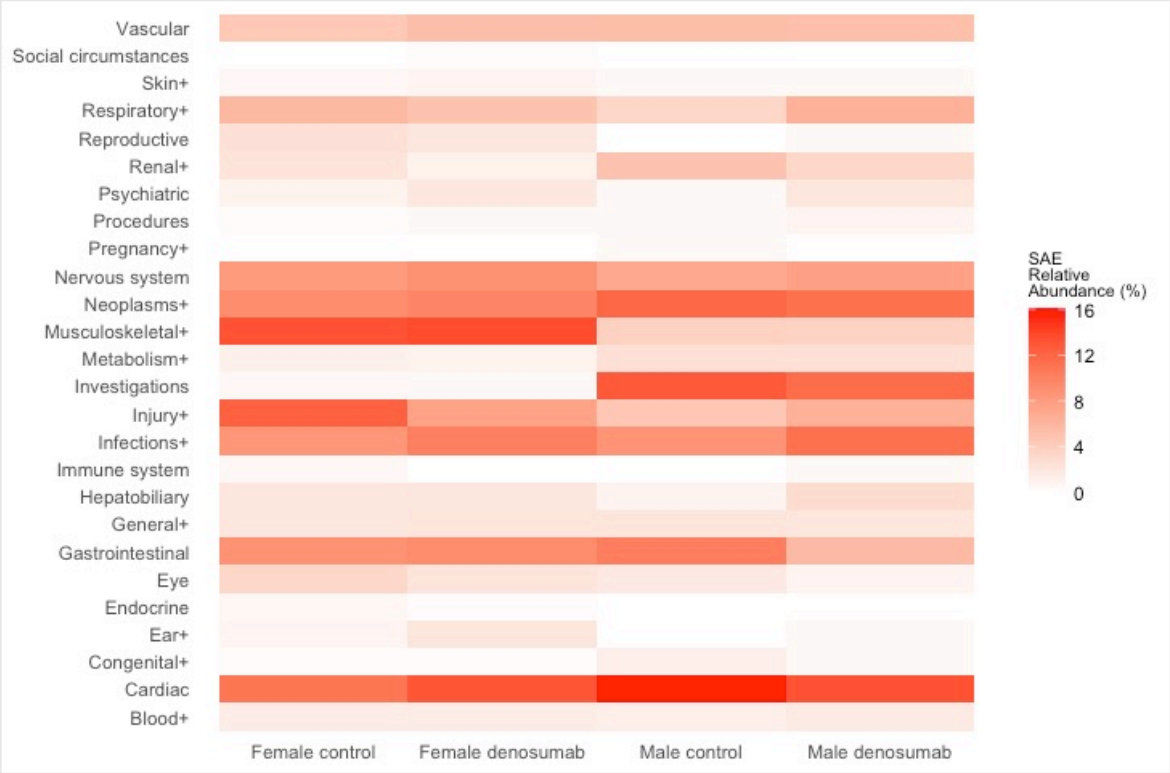

Figure 12. Heatmap of the relative abundance of a SOC, based on the serious adverse incidence in that SOC, relative to the total sum of SAE subject incidence in all SOCs, per sex, per treatment arm. SAE = Serious Adverse Event. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

Advantages
Clearly shows each individual system organ class
Easy to pick out 'outliers' (groups in which the serious adverse event subject incidence is much higher compared to the SOCs around them will stand out due to their colour intensity)
Aesthetics are pleasing and clear
Disadvantages
Hard to compare colour gradients between two to four rectangles
Relative abundances do not show absolute values and don't indicate group sizes, while it is of value to know that the female subset is much larger than the male subset of the study population
No indication of the variation of the data (adding in error bars with standard error is possible, but I assume the data is not normally distributed)
Plot does not show if there is a significant difference between the groups
The plots miss a dimension, because we want to compare both control vs. treatment and men vs. women, preferably at the same time
Detection of (sex-specific) safety signals is very subjective

B5 Stacked bar chart

The stacked bar chart consists of 4 bars that represent 4 groups: males and females in the control and denosumab treatment arm. The height of each bar is 100%, referring to the sum of all subjects with a serious adverse event, counted per system organ class. The number of subjects with an SAE per SOC is displayed as a percentage of the total number of subjects with an SAE (fig. 13). This results in twenty-six colours per bar to map to all 26 SOCs.

To tone down the number of different colours in the chart and to make SOC distinguishment easier, all SOCs with a relative abundance below 5% are pooled together and labelled as 'other'. This results in a chart with 10 SOCs, excluding 'Other'. Safety signals could be detected by comparing the size of a SOC between the four bars. If the size ratio in the female subset of control vs. denosumab differs from the ratio in the male subset, this could be seen as a sex-specific safety signal. Based on the subjective analysis of this chart, safety signals were detected in four system organ classes: gastrointestinal, injury, vascular and respiratory.

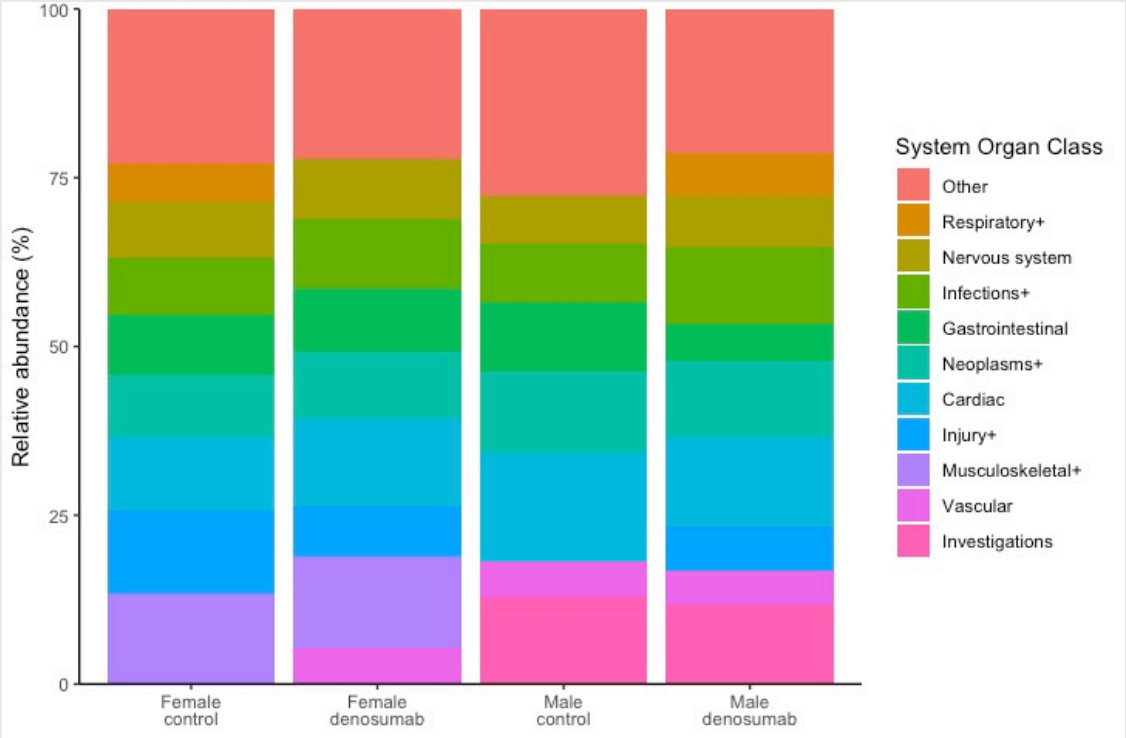

Figure 13. Stacked bar chart of the relative abundance of SOCs with a relative abundance > 5%. Based on the serious adverse incidence in that SOC, relative to the total sum of SAE subject incidence in all SOCs, per sex, per treatment arm. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

The relative abundance does not consider the size of the total treatment group, meaning the participants who did not experience an SAE (like the presentation of the heatmap). To take this into account the relative abundance can be changed to absolute abondance: the proportion of subjects per SOC based on the complete treatment arm. This results in a bar chart with varying bar heights. This presented information is still misleading, as the height of the bar does not correctly present the percentage of subjects with an SAE. This is caused by the potential double counting of subjects, who suffered from more than 1 SAE.

A last option is to not use the function to stack the bars. This will result in 26 SOCs presented next to each other. Each SOC will include 4 bars for each group (fig. 14). A static plot with this many bars is impossible to read. What could help is to make this an interactive plot. This makes the data that is

presented highly customizable by the viewer. E.g., information used to calculate the percentage of each SOC can be seen when clicking on the corresponding area in the plot. In conclusion, literature(56) uses the regular bar chart to display demographic data (both continuous and categorical variables) and to display summaries of subject exposure. The charts focus on screening and signal detection of not pre-specified AEs. This is in line with the use of the bar chart in this research.

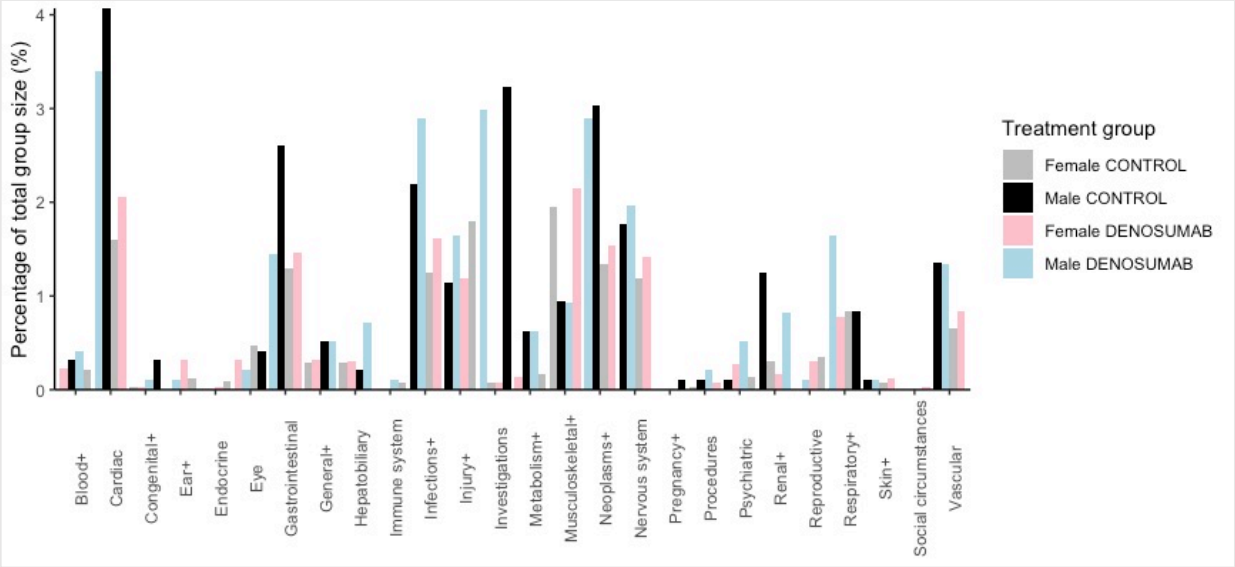

Figure 14. Bar chart of the absolute subject incidence of SAEs per treatment arm per sex. Names of system organ classes are abbreviated. Full names can be found in Appendix C1.

Advantages
Shows a clear picture of the overall pattern of relative SOC abundance
Aesthetically pleasing because there are only 4 bars
Implementation is very easy
Big differences between SOCs with high incidence are easy to spot
Many options for the presentation of data and organisation of the stacked (and grouped) bar charts
An option is to make the bar chart an interactive plot. This enables you to hover the pointer over a tile and the data is presented in a pop-up box (presented info is very customizable)
The interactive version of the grouped bar chart enables zooming in on SOC's of interest
Disadvantages
Relative abundance stacked bar charts do not show how many subjects are represented by a bar, this could make interpretation of the visual summary misleading (e.g., the female groups are much larger than male groups)
the colour shades are too close together to easily distinguish different SOCs
There is no common baseline (0 value) from which you can compare the SOCs in different bars which makes to only easily comparable SOCs the ones at the top and the bottom of the bar
SOCs that rarely occur cannot be assessed because are buried under the bigger SOCs/colours. Sex differences in these SOCs could be overlooked
No indication of the variation of the data. This issue could be solved by adding in error bars. however, this only works if the data is normally distributed, and we expect that our data is not.
Detection of (sex-specific) safety signals is very subjective

Appendix C statistical analysis

C1 Multivariate logistic regression analysis

Tsvetov2020. Denosumab-induced hypocalcaemia in patients with osteoporosis: can you know who will get low? (53)

Logistic regression is a statistical method that models the probability of a binary event taking place. It predicts the outcome of a binary variable, by analysing its relationship to one or more independent variables, which can be continuous or binary. When modelling this into a graph, an s-shaped curve will form with the y-axis depicting the probability, which can range from 0 to 1. However, when plotting logistic regression results the y-axis is often transformed to the log(odds) so the axis can range from -infinity to +infinity, transforming the s-shaped line into a linear line, allowing an easier prediction model.

In this study, multivariate logistic regression analysis is performed with treatment (control or denosumab) and sex (male of female) as variables. The coefficients give an indication in the predictive value of the variables on the odds of having a serious adverse event. Odds are calculated for the base condition, in this case female sex and control treatment. For the other conditions, odds ratios are calculated compared to the base condition. The other conditions are female sex + denosumab treatment, male sex + control treatment and male sex + denosumab treatment. The With the odds ratios (Intercept) a 95% confidence interval and p-value for significance of the predictive value of the variable are given (table 6). If the base condition is not significant this means that in this dataset the odds of a female having an SAE in the control arm is not significantly different from the odds of a female not having an SAE. The 95% confidence interval provides an estimate of the precision of the odds ratio. It describes a 95% probability that the true population value is within the odds ratio observed within your sample data. With logistic regression, we estimate the mean of the data, and the variance is derived from the mean (instead of variance estimated from the data). Since we are not estimating variance from the data, it is possible that the variance is underestimated and that the odds ratio exaggerates the size of the effect compared to a relative risk.

A significant p-value (≤ 0.05) signifies the detection threshold for safety signals. These signals can either be a treatment effect, a sex effect or an interaction between a sex and treatment effect. The last signal would indicate a sex difference in the incidence in serious adverse events attributable to denosumab treatment.

Table 6. Multivariate logistic regression analysis results. Odds (ratios) of having a serious adverse event, with 95% confidence interval and p-value. Inf = infinite, NA = Not Available.

	Odds (ratio)	**95% confidence interval**	**p**
Total subject count with an SAE			
(Base condition: sex[female] x treatment[control])	0,13	0,11 – 0,14	**<0,001**
Treatment[denosumab]	1,01	0,89 – 1,16	0,856
sex[male]	1,36	1,10 – 1,66	**0,003**
Treatment[denosumab] x sex[male]	1	0,75 – 1,33	0,995
Cardiac disorders			

(Base condition: sex[female] x treatment[control])	0,02	0,01 – 0,02	**<0,001**
Treatment[denosumab]	1,29	0,94 – 1,78	0,115
Sex[Male]	2,62	1,74 – 3,89	**<0,001**
Treatment[denosumab] * Sex[Male]	0,64	0,36 – 1,13	0,128
Musculoskeletal and connective tissue disorders			
(Base condition: sex[female] x treatment[control])	0,02	0,02 – 0,02	**<0,001**
Treatment[denosumab]	1,1	0,82 – 1,49	0,516
Sex[Male]	0,48	0,22 – 0,90	**0,036**
Treatment[denosumab] * Sex[Male]	0,9	0,33 – 2,41	0,824
Neoplasms benign, malignant and unspecified (incl. cysts and polyps)			
(Base condition: sex[female] x treatment[control])	0,01	0,01 – 0,02	**<0,001**
Treatment[denosumab]	1,15	0,81 – 1,65	0,432
sex[male]	2,31	1,45 – 3,59	**<0,001**
Treatment[denosumab] x sex[male]	0,83	0,44 – 1,56	0,557
Infections and Infestations			
(Base condition: sex[female] x treatment[control])	0,01	0,01 – 0,02	**<0,001**
Treatment[denosumab]	1,3	0,91 – 1,87	0,154
sex[male]	1,78	1,05 – 2,92	**0,026**
Treatment[denosumab] x sex[male]	1,02	0,52 – 2,02	0,95
Gastrointestinal disorders			
(Base condition: sex[female] x treatment[control])	0,01	0,01 – 0,02	**<0,001**
Treatment[denosumab]	1,14	0,79 – 1,65	0,477
sex[male]	2,05	1,25 – 3,27	**0,003**
Treatment[denosumab] x sex[male]	0,48	0,22 – 1,01	0,056
Nervous system disorders			
(Base condition: sex[female] x treatment[control])	0,01	0,01 – 0,02	**<0,001**
Treatment[denosumab]	1,19	0,82 – 1,74	0,357
sex[male]	1,49	0,83 – 2,54	0,155
Treatment[denosumab] x sex[male]	0,93	0,43 – 2,00	0,848
Injury, poisoning and procedural complications			

(Base condition: sex[female] x treatment[control])	0,02	0,01 – 0,02	**<0,001**
Treatment[denosumab]	0,65	0,46 – 0,93	**0,02**
sex[male]	0,63	0,32 – 1,14	0,157
Treatment[denosumab] x sex[male]	2,21	0,95 – 5,29	0,068
Respiratory, thoracic and mediastinal disorders			
(Base condition: sex[female] x treatment[control])	0,01	0,01 – 0,01	**<0,001**
Treatment[denosumab]	0,91	0,56 – 1,46	0,7
sex[male]	0,99	0,43 – 2,03	0,979
Treatment[denosumab] x sex[male]	2,19	0,84 – 6,05	0,116
Vascular disorders			
(Base condition: sex[female] x treatment[control])	0,01	0,00 – 0,01	**<0,001**
Treatment[denosumab]	1,28	0,78 – 2,12	0,328
sex[male]	2,08	1,04 – 3,95	**0,03**
Treatment[denosumab] x sex[male]	0,77	0,31 – 1,94	0,581
Eye disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,01	**<0,001**
Treatment[denosumab]	0,7	0,34 – 1,37	0,298
sex[male]	0,89	0,26 – 2,36	0,832
Treatment[denosumab] x sex[male]	0,71	0,09 – 4,19	0,714
General disorders and administration site conditions			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	1,16	0,54 – 2,56	0,705
sex[male]	1,86	0,59 – 5,03	0,245
Treatment[denosumab] x sex[male]	0,85	0,19 – 3,77	0,83
Reproductive system and breast disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,01	**<0,001**
Treatment[denosumab]	0,86	0,40 – 1,82	0,694
sex[male]	0	NA -Inf	0,998
Treatment[denosumab] x sex[male]	inf	0,00 – NA	0,998
Hepatobiliary disorders			

(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	1,08	0,49 – 2,40	0,852
sex[male]	0,74	0,12 – 2,73	0,696
Treatment[denosumab] x sex[male]	3,23	0,63 – 24,70	0,192
Ear and labyrinth disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	2,79	1,07 – 8,64	**0,049**
sex[male]	0	NA – Inf	0,997
Treatment[denosumab] x sex[male]	inf	0,00 – NA	0,997
Renal and urinary disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,01	**<0,001**
Treatment[denosumab]	0,53	0,20 – 1,31	0,182
sex[male]	4,15	1,86 – 9,18	**<0,001**
Treatment[denosumab] x sex[male]	1,23	0,34 – 4,54	0,755
Blood and lymphatic system disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	1,11	0,45 – 2,78	0,828
sex[male]	1,49	0,33 – 4,99	0,553
Treatment[denosumab] x sex[male]	1,19	0,21 – 7,51	0,843
Metabolism and nutrition disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	0,85	0,27 – 2,57	0,774
sex[male]	3,84	1,23 – 11,57	**0,016**
Treatment[denosumab] x sex[male]	1,16	0,24 – 5,76	0,853
Psychiatric disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	1,99	0,77 – 5,73	0,169
sex[male]	0,74	0,04 – 4,35	0,783

Treatment[denosumab] x sex[male]	2,49	0,30 – 53,58	0,449
Skin and subcutaneous tissue disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	1,66	0,41 – 8,09	0,489
sex[male]	1,49	0,07 – 11,61	0,732
Treatment[denosumab] x sex[male]	0,6	0,02 – 19,12	0,745
Endocrine disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	0,25	0,01 – 1,68	0,213
sex[male]	0	NA – Inf	0,998
Treatment[denosumab] x sex[male]	3,96	0,00 – Inf	1.000
Investigations			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	0,99	0,18 – 5,38	0,995
sex[male]	47,53	16,94 – 198,32	**<0,001**
Treatment[denosumab] x sex[male]	0,93	0,16 – 5,38	0,93
Immune system disorders			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	0	NA – Inf	0,998
sex[male]	0	NA – Inf	0,998
Treatment[denosumab] x sex[male]	inf	Inf – Inf	0,997
Social circumstances			
(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	0,999
Treatment[denosumab]	6263067024	0,00 – NA	1.000
sex[male]	2,34	0,00 – NA	1.000
Treatment[denosumab] x sex[male]	0	0,00 – Inf	1.000
Surgical and medical procedures			

(Base condition: sex[female] x treatment[control])	0	0,00 – 0,00	**<0,001**
Treatment[denosumab]	2,98	0,38 – 60,35	0,344
sex[male]	4,46	0,18 – 112,79	0,291
Treatment[denosumab] x sex[male]	0,66	0,02 – 24,28	0,807
Congenital, familiar and genetic disorders			
(Base condition: sex[female] x treatment[control])	0,01	0,01 – 0,01	**<0,001**
Treatment[denosumab]	1,05	0,97 – 1,14	0,24
sex[male]	1,55	1,38 – 1,75	**<0,001**
Treatment[denosumab] x sex[male]	0,97	0,82 – 1,14	0,676

C2 Meta-analysis

Meta-analysis combines, summarizes and interprets available evidence in a quantitative way, across multiple studies. To perform a meta-analysis an effect size must be found to be summarized across all studies. This is a metric quantifying the relationship between two treatments. It captures the direction and magnitude of this relationship. In our study this relationship is the relative risk of a serious adverse event. The meta-analysis per system organ class is visualized in a forest plot and presented and discussed in appendix B2.

We performed meta-analysis on the safety data from selected trials (table 7). The relative risk (RR) of an SAE in the denosumab vs. control treatment was calculated as the effect size. Saag et al. 2018 is split into two groups: the men and the women.

Table 7. meta-analysis of subject incidences of serious adverse events per study. Relative risks of control vs. denosumab treatment, with confidence intervals and the weight of each relative risk. Number of studies combined: k = 6. Number of observations: o = 10485. Number of events: e = 1247.

Study	**Relative risk**	**[95%-Confidence Interval]**	**%Weight(random)**
Cummings2009	0.9998	[0.8803; 1.1355]	66.8
Smith2009	1.0188	[0.8002; 1.2973]	18.6
Orwoll2012	1.1000	[0.4854; 2.4929]	1.6
Gnant2015	1.6068	[0.7981; 3.2347]	2.2
Saag2018 (Female)	0.9601	[0.6564; 1.4042]	7.5
Saag2018 (Male)	0.9097	[0.5132; 1.6126]	3.3

We assume that all calculated RRs are estimators of the true relative risk. When we pool the effects, we give RRs with a higher precision (smaller standard error) a greater weight. The Mantel-Haenszel method calculates the weights of studies with binary outcome data. This method uses the number of

events and non-events in the treatment and control group to determine a study's weight. We see that the bigger the study population is, the greater the weight attributed to the RR calculated from the trial. This makes sense, as you would expect to count more events in a bigger group. Gnant et al. 2015 is the only trial with a relatively higher RR calculated (1.6), compared to the other trials that show a relative risk close to 1, indicating that there is no risk difference between the control and treatment group. The study population of Gnant et al. consists of women with non-metastatic breast cancer receiving aromatase inhibitor therapy, which can cause secondary osteoporosis. These results could mean that this specific population suffers from more SAEs, but as not all trials with a female population show an increased risk in the denosumab treatment, we conclude this analysis does not show a sex- difference.

The random-effects model estimates the mean of the distribution of true effect sizes, in this case the mean distribution of the RRs. The model assumes that there is a distribution of true effect sizes instead of one true effect size (opposite of the fixed-effects model). This model calculated a mean RR close to 1, meaning that there is no increased risk of an SAE in the total population (males and females combined) (table 8).

Table 8. The random-effects model. CI = confidence interval. D.F = Degrees of Freedom.

Weighted average Relative Risk	**[95%-CI]**	**p-value of RR**	**Tau^2**	**I^2**	**Q**	**D.F**	**p-value of Tau^2**
1.0092	[0.9266; 1.0992]	0.7935	0 [0.0000; 0.1433]	0.0% [0.0%; 74.6%]	1.96	5	0.8548

Knapp-Hartung adjustments (Knapp & Hartung, 2003) were used to calculate the confidence interval around the pooled effect. The adjustment aims to control for the uncertainty in the estimate of the between-study heterogeneity and is based on a t-distribution, whereas significance tests for the pooled effect usually assume a normal distribution. Usually, it causes the confidence intervals to become slightly larger. Unfortunately, the random effects model cannot be applied to our way of meta-analysis, where we determine relative risks per system organ class, as the subjects in the random-effects model are counted incorrectly (as seem in appendix B2).

Meta-analysis also allows quantification of between-study heterogeneity. Tau^2 (τ^2) represents the variance of the distribution of true effect sizes (in our case the variance of true relative risks). It estimates the between-study heterogeneity. $\tau^2=0$ means that there is as good as no variance between the relative risks of the studies. Here we used the Paule-Mandel estimator to determine Tau^2. It is an ongoing research question which Tau^2 estimator performs best for different kinds of data. Which method works better depends on various parameters such as the number of studies, the number of participants in each study, the variance in study population sizes, and how big Tau^2 is. An overview paper by Veroniki et al. 2016 recommended the Paule-Mandel method for both binary and continuous effect size data(57). The I^2 Higgins and Thompson's statistic also quantifies between-study heterogeneity and is defined as the percentage of variability in the effect sized that is not caused by sampling error(58). 0% means that zero percent of the variation in relative risks is due to between-study heterogeneity. Cochran's Q is used to test if the variation in the data is what can be expected based on sampling error alone. The p-value is not below the critical 5%, therefore we can assume that the differences in RR are caused by sampling errors and less/not by between-study heterogeneity.

In subgroup analysis we assume that studies in our meta-analysis do not stem from one overall population. It is hypothesized that they fall into different subgroups and that each subgroup has its own true overall effect. We decided against subgroup analysis, because our complete safety analysis set consists of only 5 studies. When the number of studies in a subgroup is small (< 5), it is likely that the estimates will be imprecise(59).

C3 Fisher's exact test

Fisher's exact test performs a hypothesis test to determine whether there is a significant association between two categorical variables. This already implies it cannot be used to assess multiple SOCs at one, or the association between SAE incidence, treatment and sex at the same time.

We used the test to determine the significance of independence between SAEs and treatment. The incidences of serious adverse events in each treatment arm, per sex per SOC are used as input and must be filtered into a 2x2 table for the test. To determine the relative risk and Fisher's exact p-value for each SOC and in both sexes, 54 2x2 tables had to be made and tested separately. The test compares the number of 'successes' (the number of subjects who had an SAE) and 'failures' (subjects without an SAE) between the two treatment arms and calculates the significance of this association by determining the p-value (table 9). Implementation is therefore quite laborious, especially compared to the meta-analysis, which produces the same results in a single analysis. The comprehension is straightforward and cleanly visualized in the volcano plot. It lacks the direct comparison between the sexes, so sex-specific safety signals are detected by searching the p-value in both sexes. If the significances of the p-value differ, this could suggest a sex difference.

*Table 9. Fisher's exact test on serious adverse event incidence in control vs. denosumab treatment per system organ class per sex. Female subset of safety analysis set: control arm: N = 4257, denosumab arm: N = 4191. Male subset of safety analysis set: control arm N = 958, denosumab arm = 969. P.signif = significance of p-value, NS = not significant, * = significant. Inf = infinite, NA = Not Available.*

Sex	SOC	control SAE subject incidence	denosumab subject incidence	n	p	p.signif	Relative risk
Female	Total subject count with an SAE	478	486	8558	0,864	ns	1,01
Male	Total subject count with an SAE	140	143	1927	0,949	ns	1,01
Female	Cardiac	68	88	8558	0,125	ns	1,29
Male	Cardiac	39	33	1927	0,472	ns	0,84
Female	Musculoskeletal +	83	92	8558	0,542	ns	1,10
Male	Musculoskeletal +	9	9	1927	1,000	ns	0,99
Female	Neoplasms+	57	66	8558	0,468	ns	1,15
Male	Neoplasms+	29	28	1927	0,894	ns	0,95
Female	Infections+	53	69	8558	0,171	ns	1,29
Male	Infections+	21	28	1927	0,386	ns	1,32
Female	Gastrointestinal	55	63	8558	0,517	ns	1,14
Male	Gastrointestinal	25	14	1927	0,076	ns	0,55

Female	Nervous system	51	61	8558	0,392	ns	1,19
Male	Nervous system	17	19	1927	0,867	ns	1,10
Female	Injury+	77	51	8558	0,020	*	0,66
Male	Injury+	11	16	1927	0,439	ns	1,44
Female	Respiratory+	36	33	8558	0,718	ns	0,91
Male	Respiratory+	8	16	1927	0,149	ns	1,98
Female	Vascular	28	36	8558	0,380	ns	1,28
Male	Vascular	13	13	1927	1,000	ns	0,99
Female	Eye	20	14	8558	0,308	ns	0,70
Male	Eye	4	2	1927	0,450	ns	0,49
Female	General+	12	14	8558	0,845	ns	1,16
Male	General+	5	5	1927	1,000	ns	0,99
Female	Reproductive	15	13	8558	0,710	ns	0,86
Male	Reproductive	0	1	1927	1,000	ns	Inf
Female	Hepatobiliary	12	13	8558	1,000	ns	1,08
Male	Hepatobiliary	2	7	1927	0,179	ns	3,46
Female	Ear+	5	14	8558	1,000	ns	2,78
Male	Ear+	0	1	1927	0,372	ns	Inf
Female	Renal+	13	7	8558	0,188	ns	0,54
Male	Renal+	12	8	1927	0,378	ns	0,66
Female	Blood+	9	10	8558	1,000	ns	1,10
Male	Blood+	3	4	1927	1,000	ns	1,32
Female	Metabolism+	7	6	8558	0,790	ns	0,85
Male	Metabolism+	6	6	1927	1,000	ns	0,99
Female	Psychiatric	6	12	8558	0,237	ns	1,99
Male	Psychiatric	1	5	1927	0,218	ns	4,94
Female	Skin+	3	5	8558	0,726	ns	1,66
Male	Skin+	1	1	1927	1,000	ns	0,99
Female	Endocrine	4	1	8558	0,217	ns	0,25
Male	Endocrine	0	0	1927	1,000	ns	
Female	Investigations	3	3	8558	1,000	ns	0,99
Male	Investigations	31	29	1927	0,794	ns	0,92
Female	Immune system	3	0	8558	0,124	ns	0,00
Male	Immune system	0	1	1927	1,000	ns	Inf
Female	Social circumstances	0	1	8558	1,000	ns	Inf
Male	Social circumstances	0	0	1927	1,000	ns	NA
Female	Procedures	1	3	8558	0,625	ns	2,98
Male	Procedures	1	2	1927	1,000	ns	1,98
Female	Congenital+	1	1	8558	1,000	ns	0,99
Male	Congenital+	3	1	1927	0,372	ns	0,33
Female	Pregnancy+	0	0	8558	1,000	ns	NA
Male	Pregnancy+	1	0	1927	0,497	ns	0,00