Point-By-Point

Reviewer 1

Comments:

1.1) *Over the counter NSAIDs are mentioned on manuscript page 13, and that led me to wonder if the NSAIDs tracked in the study were all prescription strength, or if they varied in strength from prescription to OTC. If the increased HF risk applies only to prescription strength NSAIDs, then knowledge of this risk needs to focus on providers - if the increased risk is the same if taking OTC NSAIDs, then there is a larger public health education campaign that needs to occur.*

> ***Reply:*** We thank the Reviewer for pointing out this issue. Indeed, OTC NSAIDs are probably typically used at lower doses than prescription NSAIDs (*Duong M et al., Br J Clin Pharmacol 2013; 77:887*) but there is a potential risk for inappropriate overuse of these drugs by patients (*Koffeman AR et al. Br J Gen Pract 2014; 64:e191*). We have expanded the Discussion session of the manuscript (page 14, lines 3-9) to address these issues. Two reference were also added (References [34] and [35]).

1.2) *I would have liked to have seen gender differences reported in the results - e.g. men are at greater for risk for HF if prescribed diclofenac, while women are at greater risk if prescribed indomethacin.*

> ***Reply:*** We have added some results on gender differences in the Results section (page 11, lines 1-5). Specifically, in a stratified analysis, a statistically significant difference in the estimated ORs was observed across genders only for nimesulide, etoricoxib, and indometacin (**Table 3**).

1.3) *In the results section of the abstract, the "estimated adjusted odds ratio of...of HR hospitalization associated with current use of any NSAID was 1.20" - is that 1.20 a decimal expression of the ratio? Should it be written as a ratio instead -12:10? Or change to percent increased risk (roughly 20%)?*

Reply: We have modified the abstract as suggested by the Reviewer (page 2, lines 12-14).

Reviewer 2

Comments:

2.1) *Methods: Because the comparison group here are non-users of these medications, it is possible that patients using NSAIDs or COX-2 inhibitors may be more likely to be hospitalized due to increased exposure to their provider and the healthcare system. Is it possible to identify an control group (e.g., acetaminophen or aspirin users) that is not thought to be associated with an increased risk of HF to see whether the associations persist? Comparisons without an active comparator group may be tricky due to inherent differences in the study populations.*

Reply: We thank the Reviewer for raising this issue. Please note that, following a new-users design, all members of the study cohort must have received a dispensation of NSAIDs at the moment of cohort entry (page 5, lines 18-21). Non-users of NSAIDs were thus not included in the cohort. Accordingly, in our study, the reference group was past users of any NSAIDs, rather than non-users of NSAIDs (page 7, lines 11-14).

Still, we agree with the Reviewer that using as reference group the current users of drugs such as acetaminophen or aspirin could have improved the clinical relevance of the estimated associations. Unfortunately, acetaminophen and aspirin used as antipyretic and for pain-management are not available in the databases participating in this study since they are non-reimbursable. Additionally, there is some evidence in the literature that aspirin may increase the risk of HF (*Massie BM. J Am Coll Cardiol 2005; 46:963-6).* The choice of other possible active comparators may be affected by some degree of arbitrariness (which drug to choose?) and it may reduce power to detect any association if the chosen drug was not frequently used.

Using past users of any NSAID as the reference group circumvents both these problems. This provides a common and sizeable comparator group for all studied NSAIDs. Additionally, note that it is possible to assess the association with HF risk and current use of NSAID A in comparison with current use of NSAID B by directly comparing the estimated odds ratios for current use of A vs. past use of any NSAID and current use of B vs. past use of any NSAID.

2.2) *Methods: Please provide more information on the control group. Did they ever experience a HF hospitalization? Or just not in that period of time that they were matched?*

Reply: As now clarified in the Methods section: (a) cohort members were excluded if they had experienced a hospitalization with a primary diagnosis of HF in the year before cohort entry (page 6, point iv on lines 3-6); (b) each remaining cohort member accumulated person-years of follow-up from the date of cohort entry until the first among the dates of hospitalization with a primary diagnosis of HF (cases) of censoring (death/emigration/end of study) (page 6, line 7-11); (c) each case identified during follow-up was matched to up to 100 controls selected among all individuals who were still at risk of experiencing a hospitalization with a primary diagnosis of HF at the date when the corresponding case had experienced it (index date) (page 7, line 2-5). This means that controls must not have experienced a hospitalization with primary diagnosis of HF from one year before cohort entry to the index date of the corresponding case. However, controls may have received a secondary hospital or outpatient diagnosis of HF at any time during this period. This possibility was accounted for by covariate adjustment (page 7, line 16-22 and page 8 and lines 2-10).

2.3) *Methods: While mentioned in the discussion, please clarify more that the patients included in the study could have diagnosed HF at baseline. The motivation for the study appears to be the association between NSAID/COX-2 use and clinically-significant HF (defined here as 'hospitalization for HF'); however, if HF differs substantially between the case and control groups at baseline, naturally there would be differences in hospitalizations for HF (and it may not be* *possible to fully control for this). Why not also examine the association among patients without any HF at baseline? This information might help provide information to answer the key biological questions presented in the introduction.*

Reply: We thank the reviewer for this suggestion. In the revised version of the paper, we have provided some results on the association between current use of individual NSAIDs and the risk of HF among patients without any prior secondary hospital or outpatient HF diagnoses in the Results section (page 11, lines 1-5, and **Table 3**). In particular, a statistically significant increased risk of HF also among patients without secondary hospital or outpatient HF diagnoses was confirmed for current use of indomethacin (OR; 95% CI: 1.52; 1.31, 1.77), diclofenac (1.21; 1.16, 1.26), ketorolac (1.94; 1.71, 2.19), piroxicam (1.31; 1.21, 1.41), ibuprofen (1.15; 1.08, 1.21), naproxen (1.19; 1.08, 1.31), rofecoxib (1.34; 1.25, 1.44), etoricoxib (1.55; 1.42, 1.69), and nimesulide (1.21; 1.16, 1.27).

Additionally, we have clarified in the Methods section (page 6, lines 5-6) that patients included in the study could have diagnosed HF at baseline as per the Reviewer's suggestion. Note that, as now specified in the Methods section (page 7, line 16-22 and page 8 and lines 2-10), we adjusted our main analyses for prior history of secondary hospital or outpatient diagnoses of heart failure to account for differences in baseline HF risk among the case and control group.

Minor comments:

2.4) *Methods: Is it possible to capture previous use of NSAIDs or COX-2 inhibitors (and the extent to which that may affect the 'control' population)?*

Reply: To be included in the study, members of the cohort must not have received NSAIDs dispensations in the year preceding cohort entry, as prescribed by our inclusion criteria (page

5, point (ii) from the last line). The exclusion of prevalent NSAIDs users, i.e. patients who used NSAIDs previous to cohort entry, helped avoid several types of biases and ameliorated confounding by indication (*Ray WA. Am J Epidemiol 2003; 158:915;* this reference is now also cited in the Methods section for clarification on page 5, line 18).

2.5) *Methods: To help the reader interpret the study findings, please provide information per medication (in the Appendix) about what was considered to be "low", "medium", "high" and "very high" dose class.*

Reply: To comply with the Reviewer's suggestion, we added in **Supplementary Table S5** the information on the daily amount of active principle corresponding to each dose category for each individual NSAID considered in the dose-response analysis (page 12, lines 2-5).

2.6) *Discussion: Please provide a biological rationale here for why NSAIDs might be associated with increased risk of a HF hospitalization.*

Reply: We modified the Discussion as suggested by the Reviewer (page 12, from line 10).

2.7) *Discussion: Is there any information on the extent to which any dispensations of NSAIDs or COX-2 inhibitors may not be captured within the healthcare databases? In the US, because traditional NSAIDs are usually (and most often) dispensed without a prescription (over-the-counter), so they are frequently not captured in databases. If there are medications that are not captured in the databases, this could potentially affect the study results through misclassification of exposure, but agree that this would likely bias towards the null.*

Reply: Data from the literature suggests that, much like in the USA, the use of OTC NSAIDs in Europe is frequent. For instance, in France, *Duong M et al.* (*Br J Clin Pharmacol 2013; 77:887*) estimated that about 44% of individuals registered in the French national healthcare insurance system received at least one dispensation of NSAIDs over a two-year period. Of these, about 19% received only OTC NSAIDS, 53% received only prescription NSAIDs, while 28% received both OTC and prescription NSAIDs. In the Netherlands, likewise, *Koffeman AR et al.* (*Br J Gen Pract 2014; 64:e191*) estimated that about 30% of the general population had used OTC NSAIDs during a one-month period. More generally, in a large-scale population survey conducted in 15 European countries and Israel, *Breivik H et al.* (*Eur J Pain 2006; 10:287*) found that about 55% of individuals self-describing as chronic pain sufferers reported taking OTC NSAIDs, with prevalence ranging from 13% in Denmark and Norway to 91% in Finland.

Since the healthcare databases participating in this study only capture dispensations of prescribed NSAIDs and not dispensations of OTC drugs, these results suggest that the proportion of missed NSAID exposure may be high and possibly different across countries. We recognize that this is a potential limitation of our study, as indicated in the Discussion section of the manuscript (page 14, from line 17). Regardless, we agree with the Reviewer that the effect of exposure misclassification due to missed OTC dispensations is to lead to an underestimation of the associations of interest. We argue this point in the Discussion section of the Manuscript (page 14, lines 20-22) and also, in more detail, in our reply to Reviewer's 6 comment 3 below. Still, we also acknowledge that observed estimates may, by chance, be an overestimate (*Jurek AM, et al. Pharmacoepidemiol Drug Saf 2005; 34:680*), as we now also state in the Discussion (page 15, lines 9-11).

Discretionary comments:

2.8) *Abstract: If space permits, please provide motivation for this work here.*

Reply: We have modified the Abstract as suggested by the Reviewer (page 2, line 2).

2.9) *Abstract: Please define "current use" in the abstract as well as the cases and controls used for the comparisons.*

Reply: We have modified the Abstract as suggested by the Reviewer (page 2, line 12).

2.10) *Methods: With 27 different NSAID comparisons, please address whether there is any risk of multiple comparisons affecting the study results.*

Reply: Following *Bender & Lange (J Clin Epidemiol 2001; 54:343)*, to address the multiple comparisons issue it is important to distinguish between the *comparisonwise error rate* (CER), i.e. the Type I error rate of each tested hypothesis considered on its own, and the *experimentwise error rate* (EER), i.e. the probability of a Type I error when all tested hypotheses are considered as components of a larger single hypothesis. Importantly, *Bender & Lange* argue that

> "*if the investigator only wants to control the CER, an adjustment for multiple tests is unnecessary.*"

As we mention in the Introduction of our paper, there is a dearth of information on the risk of heart failure associated with the use of individual NSAIDs. Hence, in this study we assessed the association between 27 individual NSAIDs and the risk of heart failure. We assessed each of these associations as of interest in its own right rather that as part of a larger hypothesis. Hence the focus of this study was on individual associations and their quantification. Consequently we believe that, in this specific setting, controlling the CER should be of greater interest than controlling the EER. Therefore, in agreement with *Bender & Lange*, we did not perform multiple comparisons corrections in this study.

2.11) *Methods: Please consider whether a sensitivity analysis of any diagnosis of HF in the outpatient setting may be possible. While HF is complex, it may be worth exploring any diagnosis to provide additional generalizability to the study.*

Reply: We thank the reviewer for this suggestion. As we mention in the Methods section, we excluded outpatient HF dispensations from our outcome definition because HF involves several pathophysiological mechanisms, which, along with factors triggering circulatory decompensation, may give heterogeneous clinical manifestations. For this reason we have chosen to focus only on severe circulatory decompensations leading to an hospitalization by not including in our endpoint definition i) diagnostic codes for clinical HF in the outpatient setting and ii) secondary hospital discharge codes for HF (which likely represent HF manifestations occurring during hospitalizations for other causes) (page 6, lines 15-20). Still, in the revised version of the paper, we have provided some results on the association between current use of individual NSAIDs and the risk of HF among patients with/without any prior secondary hospital or outpatient HF diagnoses in the Results section (page 11, lines 1-5 and **Table 3**).

2.12) *Methods: Please provide a motivation for using a nested case-control study rather than a full cohort study.*

Reply: The Nested Case-Control (NCC) design is a well established alternative to the Cohort design in Epidemiology. In particular, the NCC design can yield results comparable to those of the Cohort design at a much lower computational cost, especially in large databases and when exposures are defined time-dependently (*Essebag V, et al. BMC Med Res Methodol 2005; 5:5*). Since these conditions match the setting of the present study, the NCC design was preferred over the cohort approach.

2.13) *Discussion: Was the study 'powered' sufficiently to examine all of the NSAIDs (even in a post-hoc analysis)?*

Reply: As some NSAIDs were infrequently used in the source population of this study, we expected a potentially low power to detect significant ORs for those individual NSAIDs more rarely used in the EU. This was indeed the case, for instance, for sulindac, which in our main analysis (based on pooled individual-level data from about 10 million NSAIDs users) was used by only 0.01% of controls. Using the approach of *Lui KJ (Am J Epidemiol 1988; 127:1064)*, we estimated that, with 100 controls per case, our main analysis including 92,163 HF cases could have only identified ORs of 2.00 or more with 80% power for sulindac. The pooling of individual-level data however guaranteed us a greater power to detect significant ORs for the more frequently used NSAIDs. For instance, using the same approach as above, we estimated that our main analysis had an 80% power to detect as significant ORs of 1.32 (1.17) for NSAIDs used by about 0.10% (1.00%) of controls, i.e. about as much as ketorolac (rofecoxib). Interestingly, our main analysis could have detect a significant OR associated with current use of celecoxib as low as 1.08 with an 80% power. We have added a few words in the Discussion section about this issue (page 13, lines 7-12).

2.14) *Discussion: The study appears to be potentially underpowered for some dose classes (Figure 3). Please provide this as a limitation.*

> ***Reply:*** Power was lower in the dose-response analysis because only two databases (PHARMO and THIN) were considered. This is a limitation of our study that we now mention in the Discussion (page 15, line 12-17).

Reviewer 3

Comments:

3.1) *The study cohorts consisted of new users of NSAIDs, with past users of NSAIDs (i.e., those with no use in the 183 days before the index date) as the reference group for all analyses. This reference group renders the results somewhat difficult to interpret. These past users will have discontinued use for a variety of reasons, including improved symptoms, the occurrence of side effects, and poor adherence. Did the authors consider using current users of one of the individual NSAIDs are the reference group? Such a comparison may also be more clinically relevant, particularly given the information presented in the second to last paragraph of the introduction.*

> ***Reply:*** We thank the Reviewer for raising this issue. Indeed, using as reference group the current users of one of the individual NSAIDs considered in this study could have improved the clinical relevance of the estimated associations. Still, we feel that any such choice could have been affected by some arbitrariness (which NSAID?) and it may have reduced power to detect any association if the chosen NSAID was not frequently used. Using past users of any NSAID as the reference group circumvents both these problems. This provides a common and sizeable comparator group for all studied NSAIDs. Additionally, note that it is possible to assess the association with HF risk and current use of NSAID A in comparison with current use of NSAID B by directly comparing the estimated odds ratios for current use of A vs. past use of any NSAID and current use of B vs. past use of any NSAID.

3.2) *It is also likely that some of the past users were prescribed short-term NSAID therapy for an acute event (e.g., sports injury, dental work), and these patients may have an inherently different risk of HF relative to those who are using NSAIDs long-term to treat chronic conditions. Similarly, different NSAIDs may be prescribed preferentially for different indications. Some discussion of confounding by indication is warranted.*

Reply: We recognize that patients included in this study may have been different with respect to the main indication and intended duration of NSAIDs therapy. In fact, we clarified in the Discussion section (page 15, starting from line 18) that

> *"[...] the impact of heterogeneity in baseline patients' characteristics must be considered when interpreting our findings. Indeed, some individual NSAIDs more frequently used for different acute or chronic indications could have resulted in different patterns of use, as well as in different types of populations of users*."

We believe that the adopted covariate adjustment procedure (which accounted for individual-level covariates potentially associated with contraindications for NSAIDs, e.g. prior cardiovascular diseases) and the random-effects meta-analytic approach (which accounted for potential heterogeneity in prescribing practices across DBs) should have somewhat protected our conclusions with respect to these issues. Nevertheless, we recognize that residual differences in patient's baseline characteristics may account for some of the observed variations in relative risk estimates associated with different individual NSAIDs, as we now acknowledge in the Discussion (page 16, line 5-7).

3.3) *Was a minimum duration of NSAID prescription required for cohort entry? In addition, was inclusion restricted based on route of NSAID administration? For example, could a patient enter the cohort due to a prescription of a topical NSAID?*

Reply: Duration of therapy with NSAIDs was not an inclusion/exclusion criteria for entry in the study cohort. Additionally, only NSAIDs with Anatomic-Therapeutic-Chemical (ATC) code M01* were considered in this study. This effectively excludes topical NSAIDs, which have (ATC) code M02*, as now clarified in the Methods section of the manuscript (page 5, lines 19-21).

3.4) *Controls were selected using incidence density sampling, matching on database, sex, age at cohort entry, and date of index prescription. With all patients entering the cohort on an NSAID prescription (and thus initially exposed), person-moments earlier during follow-up were thus more likely to be exposed to an NSAID than those later during follow-up. Did the authors consider also matching on duration of follow-up (via risk set sampling) to ensure that cases and controls had the same opportunity for exposure?*

Reply: We thank the Reviewer for point out this unclear passage of our paper. Indeed, as we have now clarified in the Methods section (page 7, lines 2-5), matching was performed by risk-set sampling. Specifically, for each case, controls were selected among cohort members whose follow-up did not end before the case's outcome onset date.

3.5) *In the Introduction, the authors discuss that current guidelines limit the use of NSAIDs in patients predisposed to HF and prohibit their use in patients with diagnosed HF. With approximately 9% of cases and 2.5% of controls having a history of HF, the inclusion of some subgroup analyses and/or tests for interaction would be informative.*

Reply: We have now added to the Results section of out manuscript (page 11, lines 1-5 and **Table 3**) a few detail on the performed subgroup analysis comparing the risk of HF associated with use of individual NSAIDs among patients with or without prior secondary hospital or outpatient HF diagnoses. No statistically significant NSAID-prior HF interaction was detected, although power may have been low in this subgroup analysis (page 13, line 3-6).

3.6) *In site-specific analyses, covariates should be included in the model based on substantive knowledge and not p-values. If an automated variable selection process is used, the use of AIC or BIC is preferred.*

Reply: Our covariate-selection procedure was two-pronged: first, several covariates, selected on the basis of substantive knowledge and available in all databases, were forced in the model (*a-priori covariates*); second, a set of remaining candidate covariates were included in the model if they passed a backward selection procedure with a p-value for removal of 0.10 (*candidate covariates*). The candidate covariates, initially considered for their substantive importance, were screened by the backward selection procedure because they were potentially less well captured by the participating databases. We stress again that a-priori covariates were all deemed as important potential confounders and did not enter the backward selection procedure.

Furthermore, it must be noted that, since all of the considered candidate covariates were dichotomous, the AIC selection procedure is equivalent to a backward selection procedure with a p-value for removal of 0.157 (*Steyerberg EW et al. Statist Med 2000; 19:1059*). Hence we do not expect the results of such procedure to be too different from the ones obtained from the implemented procedure, which considered a p-value for removal of 0.100. The BIC typically selects less variables than the AIC (*Burham KP, et al. Socio Meth Res 2004; 33:261*).

Additionally, we performed unadjusted analyses (data not shown in the paper) and the results did not substantially change. This suggests that covariate selection should not have heavily affected our results.

3.7) *The inclusion of duration-response analyses would also be helpful. For example, some evidence suggests that the increase in myocardial infarction with rofecoxib only occurred after 18 months of use.*

Reply: We performed a duration-response analysis on the basis of pooled individual-level data from all databases. Specifically, for all current users of NSAIDs we computed the duration of continuous use of each currently used individual NSAID by assuming a daily dose of 1 DDD. Duration of use was categorized as short (<7 days), medium (7-29 days), long (30-89 days), and very long (≥90 days). A conditional logistic regression model including the indicator variables of these categories (including only NSAIDs with at least one case in each category) was then implemented to estimate the adjusted ORs for HF hospitalization comparing each category of current-use duration with respect to past use of any NSAIDs. The results are summarized in the following table:

Individual NSAIDs	***OR***	***95% Lower CL***	***95 Upper CL***
Aceclofenac - SHORT (<7 days)	0.921	0.666	1.271
Aceclofenac - MEDIUM (7-29 days)	1.160	1.014	1.327
Aceclofenac - LONG (30-89 days)	0.581	0.394	0.857
Aceclofenac - VERY LONG (>=90 days)	0.879	0.434	1.780
Acemetacin - SHORT (<7 days)	2.840	1.006	8.019
Acemetacin - MEDIUM (7-29 days)	1.328	0.643	2.742
Acemetacin - LONG (30-89 days)	0.540	0.131	2.228
Acemetacin - VERY LONG (>=90 days)	0.907	0.216	3.814
Celecoxib - SHORT (<7 days)	1.151	0.970	1.364
Celecoxib - MEDIUM (7-29 days)	0.926	0.847	1.013
Celecoxib - LONG (30-89 days)	0.921	0.826	1.026
Celecoxib - VERY LONG (>=90 days)	0.976	0.855	1.114
Dexibuprofen - SHORT (<7 days)	1.385	0.727	2.639
Dexibuprofen - MEDIUM (7-29 days)	1.091	0.737	1.614
Dexibuprofen - LONG (30-89 days)	1.073	0.394	2.925
Dexibuprofen - VERY LONG (>=90 days)	3.544	1.252	10.032

Diclofenac - SHORT (<7 days)	1.476	1.371	1.589
Diclofenac - MEDIUM (7-29 days)	1.128	1.070	1.189
Diclofenac - LONG (30-89 days)	1.123	1.031	1.223
Diclofenac - VERY LONG (>=90 days)	1.096	0.976	1.231
Diclofenac, combi. - SHORT (<7 days)	1.140	0.859	1.512
Diclofenac, combi. - MEDIUM (7-29 days)	0.963	0.834	1.112
Diclofenac, combi. - LONG (30-89 days)	1.191	0.987	1.438
Diclofenac, combi. - VERY LONG (>=90 days)	0.915	0.739	1.132
Etodolac - SHORT (<7 days)	0.577	0.080	4.145
Etodolac - MEDIUM (7-29 days)	0.929	0.507	1.702
Etodolac - LONG (30-89 days)	1.097	0.616	1.954
Etodolac - VERY LONG (>=90 days)	0.744	0.452	1.223
Etoricoxib - SHORT (<7 days)	1.672	1.343	2.082
Etoricoxib - MEDIUM (7-29 days)	1.568	1.417	1.735
Etoricoxib - LONG (30-89 days)	1.255	1.071	1.470
Etoricoxib - VERY LONG (>=90 days)	1.390	1.141	1.695
Flurbiprofen - SHORT (<7 days)	0.874	0.387	1.972
Flurbiprofen - MEDIUM (7-29 days)	1.033	0.645	1.655
Flurbiprofen - LONG (30-89 days)	0.536	0.132	2.171
Flurbiprofen - VERY LONG (>=90 days)	1.481	0.534	4.104
Ibuprofen - SHORT (<7 days)	1.176	1.045	1.324
Ibuprofen - MEDIUM (7-29 days)	1.154	1.086	1.226
Ibuprofen - LONG (30-89 days)	1.241	1.109	1.388
Ibuprofen - VERY LONG (>=90 days)	1.188	1.023	1.379
Indometacin - SHORT (<7 days)	1.529	1.199	1.951
Indometacin - MEDIUM (7-29 days)	1.583	1.344	1.865
Indometacin - LONG (30-89 days)	1.306	0.905	1.885
Indometacin - VERY LONG (>=90 days)	1.145	0.643	2.037
Ketoprofen - SHORT (<7 days)	1.104	0.964	1.263
Ketoprofen - MEDIUM (7-29 days)	0.976	0.878	1.085
Ketoprofen - LONG (30-89 days)	1.034	0.856	1.250
Ketoprofen - VERY LONG (>=90 days)	1.199	0.901	1.594
Ketorolac - SHORT (<7 days)	1.904	1.710	2.120
Ketorolac - MEDIUM (7-29 days)	1.468	1.132	1.903
Ketorolac - LONG (30-89 days)	1.605	0.905	2.849
Ketorolac - VERY LONG (>=90 days)	2.147	0.647	7.122
Lornoxicam - SHORT (<7 days)	1.326	0.698	2.517
Lornoxicam - MEDIUM (7-29 days)	1.033	0.716	1.489
Lornoxicam - LONG (30-89 days)	0.790	0.373	1.672
Lornoxicam - VERY LONG (>=90 days)	1.785	0.545	5.846
Meloxicam - SHORT (<7 days)	1.228	0.977	1.544
Meloxicam - MEDIUM (7-29 days)	1.057	0.940	1.188
Meloxicam - LONG (30-89 days)	0.953	0.797	1.140
Meloxicam - VERY LONG (>=90 days)	0.920	0.771	1.098
Nabumetone - SHORT (<7 days)	0.421	0.104	1.699
Nabumetone - MEDIUM (7-29 days)	1.003	0.679	1.483

Nabumetone - LONG (30-89 days)	1.077	0.630	1.842
Nabumetone - VERY LONG (>=90 days)	1.654	1.082	2.527
Naproxen - SHORT (<7 days)	1.466	1.168	1.840
Naproxen - MEDIUM (7-29 days)	1.147	1.013	1.299
Naproxen - LONG (30-89 days)	1.086	0.922	1.280
Naproxen - VERY LONG (>=90 days)	1.146	0.931	1.411
Nimesulide - SHORT (<7 days)	1.301	1.185	1.428
Nimesulide - MEDIUM (7-29 days)	1.126	1.073	1.180
Nimesulide - LONG (30-89 days)	1.430	1.263	1.619
Nimesulide - VERY LONG (>=90 days)	1.503	1.120	2.015
Piroxicam - SHORT (<7 days)	1.546	1.372	1.741
Piroxicam - MEDIUM (7-29 days)	1.244	1.138	1.360
Piroxicam - LONG (30-89 days)	0.941	0.783	1.129
Piroxicam - VERY LONG (>=90 days)	1.249	0.912	1.710
Proglumetacin - SHORT (<7 days)	1.474	0.355	6.127
Proglumetacin - MEDIUM (7-29 days)	1.032	0.485	2.197
Proglumetacin - LONG (30-89 days)	0.821	0.302	2.231
Proglumetacin - VERY LONG (>=90 days)	1.034	0.326	3.284
Rofecoxib - SHORT (<7 days)	1.580	1.317	1.896
Rofecoxib - MEDIUM (7-29 days)	1.417	1.296	1.551
Rofecoxib - LONG (30-89 days)	1.303	1.170	1.452
Rofecoxib - VERY LONG (>=90 days)	1.196	1.037	1.378
Tenoxicam - SHORT (<7 days)	0.934	0.512	1.705
Tenoxicam - MEDIUM (7-29 days)	1.177	0.823	1.682
Tenoxicam - LONG (30-89 days)	0.809	0.331	1.977
Tenoxicam - VERY LONG (>=90 days)	1.088	0.346	3.418
Valdecoxib - SHORT (<7 days)	0.908	0.217	3.801
Valdecoxib - MEDIUM (7-29 days)	1.374	0.824	2.289
Valdecoxib - LONG (30-89 days)	1.343	0.842	2.141
Valdecoxib - VERY LONG (>=90 days)	0.239	0.033	1.727

Consistently with the findings of Huerta et al. (*Heart 2006; 92:1610–5*), no clear duration-response relationship was found between use of individual NSAIDs and HF risk.

Minor comments:

3.8) *Abstract, methods: Please mention how cases were matched to controls.*

Reply: We have now specified more details on the matching process in the Abstract (page 2, line 8).

3.9) *Abstract, results: Please mention that the reference group was past users of NSAIDs.*

Reply: We have modified the Abstract as suggested by the Reviewer (page 2, line 14).

3.10) *It would be helpful to mention in the "Harmonization and Data Processing" section that a common data model was used in the primary analysis.*

Reply: We have added this information as suggested by the Reviewer (page 5, lines 10-12).

3.11) *Covariates: Some of the descriptions of covariates are somewhat vague (e.g., "other drugs for CV diseases", "specific drugs as proxies for certain diseases and conditions not well recorded in the databases"). Please provide a complete list as an appendix.*

Reply: In the new version of the paper, we removed these ambiguous descriptions from the text and specified that the full list of covariates is available in **Table 2** (page 7, lines 16-22).

Reviewer 4

Comments:

4.1) *This study investigated the association between non-steroidal anti-inflammatory drugs and the risk of heart failure using data from 4 population-level hospital databases. The authors did a very thorough job. My only question is that why DB-specific results were not provided for the dose-response analysis? Why a similar verification approach as the main analysis was not used for the dose-response analysis?*

> ***Reply:*** Since the dose-response analysis could only be implemented in two databases (THIN and PHARMO), the number of cases available to perform the analyses was greatly reduced in comparison with the main analysis. Thus, due to sparse data issues, the dose-response analysis could only be sensibly performed by pooling individual-level data from both two databases.

Reviewer 5

Comments:

5.1) *I had some problems understanding the datasets used for the study and how participants entered the database's and the completeness of follow-up. This is a major issue and could influence detection of the primary outcome as well as exposure to NSAIDs. Please clarify.*

Reply: We have expanded the Methods section of the manuscript to better describe the databases (DBs) used for the study (page 4, lines 11-20). Participants entered each of the databases for different reasons (e.g. being a resident of the Italian Lombardy Region for the Italian DBs, or being enrolled in one of 4 Statutory Health Insurances for the German DB). Regardless, each DB longitudinally records healthcare data on all individuals in its specific target population.

Each of these databases had a different data availability period and patients may have left the corresponding DB for any reason (e.g. death or emigration from the geographical area covered by the DB), impairing completeness of follow-up. Nevertheless, our study design should have protected our findings with respect to this issue. First, the primary outcome was defined as the first hospitalization with a primary discharge diagnosis of HF during follow-up. Although it can be expected that this definition yielded a low sensitivity (some HF cases may have been missed because of loss to follow-up or because HF onset did not lead to an hospitalization), we expect specificity of outcome detection to be high. Thus errors in the ascertainment of the primary outcome should have had a minor impact on our findings.

Second, this study was conducted according to a nested case-control design, which directly accounts for differences in follow-up duration between patients. Specifically, for each identified HF case, controls were selected among cohort members with the same cohort entry date (± 4 weeks) and whose follow-up duration did not end before the index date (the date of the first HF hospitalization) of the case. For each case and matched controls, only exposure information in the period between cohort entry and the case's index date was required to conduct the analysis. Since data on all NSAIDs dispensations is available for this period in all DBs, completeness of follow-up should not have affected our exposure assessment.

5.2) *In the methods section you describe how you excluded individuals who did not have at least one year of uninterrupted observation prior to the index prescription date. How does this affect your prior history of e.g. heart failure or cardiovascular disease?*

Reply: We excluded individuals who did not have at least one year of uninterrupted observation prior to the index prescription date in order to guarantee enough time to assess prior history of heart failure or CV disease for each cohort member. We have now clarified this in the Methods section of the manuscript (page 5, lines 22-23).

5.3) *There is lot of heterogeneity in data between the individual databases and you describe that you harmonized data during data management and before pooling datasets. Did that in any way influence cohort selection or completeness of data?*

Reply: The data harmonization process in the SOS project followed the same procedure developed by and implemented in the European eu-ADR project (*Avillach P, et al. Stud Health Technol Inform 2009; 150:190*). Briefly, the harmonization process consisted first in the automatic projection of UMLS concepts in the terminologies of the coding system (e.g. ICD-9 codes) used by the various databases, followed by a manual review of the resulting concept by the database administrators and physicians. This allowed to build a common semantic basis across all DBs to guide data extraction procedures despite the heterogeneity between databases. Importantly, this process did not influence cohort selection or completeness of follow-up since it only aimed to guarantee that the data extraction processes implemented in each database were equivalent across the different used coding systems and terminologies (e.g. ICD-9 codes vs. ICD-10 codes vs. READ codes).

5.4) *I do not agree with your approach to only include primary discharge code of heart failure to define outcome since heart failure is often associated with other cardiovascular diagnoses, e.g. myocardial infarction or atrial fibrillation, and therefore affect detection of outcome events. Although the effect of this most likely would be non-differential, we know that NSAIDs increase risk of ischemic heart disease and atrial fibrillation which might affect how discharge codes were prioritized and therefore heart failure events missed.*

We know from other databases that the sensitivity of heart failure diagnosis is notoriously poor, but with acceptable positive predictive value and specificity. This would be even more pronounced by only including primary discharge coding diagnoses.

Not all heart failure patients are hospitalized and many are managed as outpatients in specialized heart failure clinics or outpatient clinics. This is being more frequent during the last years and therefore this might affect detection of heart failure events if you only included inpatients.

Reply: We agree with the Reviewer that sensitivity of heart failure as assessed by primary hospital discharge diagnoses may have been poor in this study because of issues in coding of hospital diagnoses. We now stated this more clearly in the discussion section of the manuscript (page 14, from line 23).

In particular, consistently with the Reviewer's expectation: i) data from a validation study in the OSSIFF database yielded a 80% positive predictive value; ii) high positive predictive values have been reported by other investigations based on healthcare databases for hospital discharge HF diagnosis codes considered in our study (see Reference 38). In addition, the incidence of almost 37.5 HF cases every 10,000 person-years observed in our study does not substantially differ from rates reported by available population-based investigations (see Reference 39) (page 14, from line 23). Therefore, we think the high specificity of outcome ascertainment obtained in this study should have protected our association estimates from the impact of potential misclassification errors. As the Reviewer's acknowledge, this should be even more pronounced as only primary discharge codes diagnoses were included.

Additionally, in agreement with the Review, we do not expect for outcome misclassification due to issues in coding of inpatient diagnoses to be differential with respect to NSAIDs use. All these considerations lead us to believe that, at most, outcome misclassification in this study should have produced a bias towards underestimation of the true NSAID-HF associations, as we acknowledge in the Discussion (page 15, lines 7-9).

We also agree with the Reviewer that many HF patients are nowadays managed in the outpatient setting. Still, as we mention in the Methods section, we excluded outpatient HF dispensations from our outcome definition because HF involves several pathophysiological mechanisms, which, along with factors triggering circulatory decompensation, may give heterogeneous clinical manifestations. For this reason we have chosen to focus only on severe circulatory decompensations leading to an hospitalization by not including in our endpoint definition i) diagnostic codes for clinical HF in the outpatient setting and ii) secondary hospital discharge codes for HF (which likely represent HF manifestations occurring during hospitalizations for other causes) (page 6, lines 15-20). Still, in the revised version of the paper, we have provided some results on the association between current use of individual NSAIDs and the risk of HF among patients with/without any prior secondary hospital or outpatient HF diagnoses in the Results section (page 11, lines 1-5 and **Table 3**).

5.7) *It is intriguing that you do not find any association between celecoxib and increased risk of heart failure hospitalization, which is further supported in the dose response analyses. During the study period, there has been lot of discussion on the unfavourable effect of NSAIDs on CV risk and in particular the selective COX-2 inhibitors. Thus, this could have directed physicians towards more use of non-selective NSAIDs and less use of celecoxib in high-risk patients. Did you see any trend towards change in use of different NSAIDs according to risk profile of patients (e.g. lower CV disease or hypertension in celecoxib users) during the study period?*

Reply: We agree with the Reviewer that, after the withdrawal of rofecoxib at the end of 2004, high-risk patients may have been directed towards use of non-selective NSAIDs. Indeed, there is some evidence that in the aftermath of rofecoxib's withdrawal, patients and physicians used newly-available information on the safety of coxibs to reduce treatment (*Valkhoff VE, et al. Aliment Pharmacol Ther 2012; 36: 790–9; Thiebaud P, et al. Value in Health 2006; 9:361-8; Alacqua M, et al. Arthritis & Rheumatism 2008; 59:568-74*).

To address this issue, we estimated the adjusted ORs measuring the association between current use of individual NSAIDs and the risk of HF (in comparison with past use of any NSAIDs) on the basis of pooled individual-level data from all databases i) censoring cases at 31 December 2004 (to only consider a period were all coxibs were on the market) and ii) restricting the analysis to patients who started NSAID therapy after 1 January 2006 (to only consider the period subsequent to the withdrawal of rofecoxib and valdecoxib). The results are summarized in the following table:

	Until 31 Dec 2004		**From 1 Jan 2006**		**Between-periods comparison**
	OR	**95%CI**	**OR**	**95%CI**	**p-value**
Current use of					
Indometacin	1,41	(1.14, 1.75)	1,59	(1.27, 1.98)	0.914
Sulindac	1,86	(0.89, 3.89)	2,05	(0.68, 6.16)	0.964
Diclofenac	1,16	(1.08, 1.24)	1,22	(1.15, 1.3)	0.953
Etodolac	1,00	(0.60, 1.66)	0,87	(0.53, 1.43)	0.855
Acemetacin	1,39	(0.12, 16.04)	1,57	(0.79, 3.12)	0.977
Proglumetacin	0,52	(0.18, 1.49)	1,14	(0.53, 2.46)	0.369

Ketorolac	2,22	(1.90, 2.60)	1,43	(1.14, 1.79)	0.748
Aceclofenac	1,03	(0.80, 1.32)	0,98	(0.82, 1.16)	0.946
Diclofenac, combinations	1,02	(0.88, 1.18)	1,06	(0.89, 1.27)	0.960
Piroxicam	1,24	(1.11, 1.38)	1,41	(1.25, 1.6)	0.894
Tenoxicam	0,94	(0.62, 1.42)	0,81	(0.41, 1.58)	0.838
Lornoxicam	2,77	(0.91, 8.43)	1,10	(0.79, 1.54)	0.707
Meloxicam	0,99	(0.86, 1.13)	0,97	(0.85, 1.12)	0.977
Ibuprofen	1,12	(1.03, 1.23)	1,20	(1.12, 1.3)	0.935
Naproxen	1,15	(1.00, 1.32)	1,25	(1.08, 1.44)	0.924
Ketoprofen	1,01	(0.88, 1.15)	1,08	(0.95, 1.22)	0.930
Flurbiprofen	1,24	(0.70, 2.18)	0,86	(0.46, 1.62)	0.686
Oxaprozin	0,85	(0.48, 1.53)	0,66	(0.33, 1.33)	0.704
Dexibuprofen	0,60	(0.08, 4.42)	1,38	(0.96, 1.98)	0.544
Celecoxib	0,98	(0.91, 1.06)	0,94	(0.82, 1.08)	0.952
Rofecoxib	1,37	(1.28, 1.46)	-	-	-
Valdecoxib	1,16	(0.69, 1.95)	-	-	-
Etoricoxib	1,74	(1.47, 2.05)	1,47	(1.32, 1.64)	0.886
Nabumetone	1,04	(0.71, 1.54)	1,42	(0.92, 2.19)	0.750
Nimesulide	1,24	(1.16, 1.32)	1,14	(1.06, 1.24)	0.922

The ORs for celecoxib and etoricoxib, two coxibs that have been on the market throughout the study period, are comparable between periods. This suggests that changes in the indications of coxibs due to the withdrawal of rofecoxib and valdecoxib should not have had a major impact on our results.

5.8) *There are some important differences in baseline characteristics between the cases and controls that could influence the results. In particular, there is much more frequent use of cardiac pharmacotherapy in general among cases compared to controls. Use of these drugs could represent risk factors for heart failure (e.g. hypertension or atrial fibrillation) that are being treated, or established heart failure that had not yet been hospitalized and managed in outpatient clinic or primary care. This if quite fundamental and could influence results, therefore the authors could have considered matching fewer controls on each case (it is bit of an overkill to match 100 controls on each case) and tried to achieve better balance in baseline characteristics between cases and controls, for example by using propensity score based matching or matching on more parameters. I would very much like to see results were*

Reply: Although there were some important differences in baseline characteristics between the cases and controls, in our analysis such baseline imbalances were taken into account by direct regression adjustment in the conditional logistic models. This should have protected our findings with respect to the impact of these known and measured potential confounders. Adjustment methods based on the propensity score, such as propensity score matching, are only able to address imbalances in known and measured confounders and therefore would not provide extra protection than our implemented approach (see, for instance, *Winkelmayer WC, Kurth T. Nephrol Dial Transplant 2004; 19:1671*). Additionally, due to data sharing and processing constraints, matching criteria based on the propensity score are not readily implemented (see the "Harmonization and data processing" paragraph).

Reviewer 6

Comments:

6.1) *The data are pooled from 5 electronic databases covering 4 European countries and were 'harmonised' to obtain comparable "…variable and outcome definitions…". What evidence is there to show that the resulting information is equivalent? Also, patients from the databases appear to form quite different cohorts. For example, 16% of cases in the Germany cohort had acute MI as a comorbidity compared to 1% in the UK cohort and 2% in the Italy cohort.*

Reply: The data harmonization process in the SOS project followed the same procedure developed by and implemented in the European eu-ADR project (*Avillach P, et al. Stud Health Technol Inform 2009; 150:190*). Briefly, the harmonization process consisted first in the automatic projection of UMLS concepts in the terminologies of the coding system (e.g. ICD-9 codes) used by the various databases, followed by a manual review of the resulting concept by the database administrators and physicians. This allowed to build a common semantic basis across all DBs to guide data extraction procedures despite the heterogeneity between databases. Importantly, this process is not meant to guarantee that the information extracted by each database is equivalent (as this depends on the characteristics of each single database), but rather that the data extraction processes implemented in each database were equivalent across the different used coding systems and terminologies (e.g. ICD-9 codes vs. ICD-10 codes vs. READ codes).

This fact is also highlighted, as the Reviewer points out, by the observed differences in the proportion of HF cases who were identified as having acute MI across countries. These differences may be in part due to differences in coding practices across databases, but they may also reflect a heterogeneity in the underlying covered populations. For instance, in 2012, the number of hospital discharges for cardiovascular disease, coronary heart disease, and cerebrovascular disease per 100,000 individuals was 3,500 in Germany, 1,291 in the UK, and 2,120 in Italy (*Nichols M, et al. Cardiovascular disease in Europe 2014: epidemiological update. Eur Heart J 2014; 7:2950*).

We have modified the Methods section of the manuscript to better clarify the role of the harmonization process in this study (page 5, lines 4-12).

6.2) *Cases were those hospitalised for HF during follow-up. But what percentage of heart failure patients are hospitalised? Does this vary between countries?*

Reply: It should be considered that sensitivity of heart failure as assessed by hospital discharge diagnoses may have been poor in this study because of missed outpatient diagnoses, as we state in the discussion section of the manuscript (page 14, from line 23). On the other hand, the high specificity of outcome ascertainment should have protected our association estimates from the impact of potential case-ascertainment errors.

Additionally, as highlighted in the Methods section (page 6, lines 13-20), HF is a clinical syndrome involving several pathophysiological mechanisms which, along with factors triggering circulatory decompensation, may give heterogeneous clinical manifestations which often receive delayed diagnosis. Therefore, our endpoint definition did not include outpatient diagnostic codes for clinical HF.

As the Reviewer suspects, the probability of being hospitalized when HF symptoms occur may vary between countries, possibly explains part of the heterogeneity in DB-specific odds ratios observed in this study.

6.3) *What is the estimated percentage of 'missed' NSAID drug exposure (due to 'over the counter' prescribing)? Does this vary between countries? If the percentage is large then I am not sure an 'understatement' of the actual association between NSAIDs and HF risk can be claimed.*

Reply: Data from the literature suggests that the use of OTC NSAIDs in Europe is frequent. For instance, in France, *Duong M et al.* (*Br J Clin Pharmacol 2013; 77:887*) estimated that about 44% of individuals registered in the French national healthcare insurance system received at least one dispensation of NSAIDs over a two-year period. Of these, about 19% received only OTC NSAIDS, 53% received only prescription NSAIDs, while 28% received both OTC and prescription NSAIDs. In the Netherlands, likewise, *Koffeman AR et al.* (*Br J Gen Pract 2014; 64:e191*) estimated that about 30% of the general population had used OTC NSAIDs during a one-month period. More generally, in a large-scale population survey conducted in 15 European countries and Israel, *Breivik H et al.* (*Eur J Pain 2006; 10:287*) found that about 55% of individuals self-describing as chronic pain sufferers reported taking OTC NSAIDs, with prevalence ranging from 13% in Denmark and Norway to 91% in Finland.

Since the healthcare databases participating in this study only capture dispensations of prescribed NSAIDs and not dispensations of OTC drugs, these results suggest that the proportion of missed NSAID exposure may be high and possibly different across countries. We recognize that this is a potential limitation of our study, as indicated in the Discussion section of the manuscript (page 14, starting from line 17).

Regardless, we believe that the effect of exposure misclassification due to missed OTC dispensations is to lead to a bias towards the null for of the associations of interest. This is because:

1) Since in our study patients were classified as current users of NSAIDs if they received prescription NSAIDs, missed dispensations of OTC NSAIDs at most resulted in patients being erroneously classified as non-current NSAIDs users. Hence, we expect that exposure assessment did not result in any false positive (i.e. non-current users of OTC or prescription NSAIDs classified as current NSAIDs users) among cases or controls.
2) Several covariates possibly associated with use of OTC NSAIDs, such as database of enrolment, age, gender, and cardiovascular comorbidities (*Duong M et al., Br J Clin Pharmacol 2013; 77:887; Delaney JA et al. Pharmacoepidemiol Drug Saf 2011; 20:83*), were directly adjusted for in the analysis (by matching or regression adjustment). Although we cannot exclude residual imbalances, these adjustments should have protected our association estimates from differences in the distribution of OTC NSAIDs use between cases and controls. In other words, we expect that the proportion of false negatives (i.e. current users of NSAIDs incorrectly classified as non-users) to be comparable between cases and controls within the strata of the considered covariates.

   Consequently, in agreement with the results of *Yood MU et al. (Pharmacoepidemiol Drug Saf 2007; 16:961*) and as stated in the Discussion section (page 14, lines 20-22), we expect that exposure misclassification due to missed OTC NSAIDs dispensations in our study to be non-differential conditionally on covariates and so to have the tendency to drag estimated associations towards the null. Still, observed estimates may by chance be an overestimate (*Jurek AM, et al. Pharmacoepidemiol Drug Saf 2005; 34:680*), as we now also acknowledge in the Discussion (page 15, lines 9-11).

6.4) *"…up to 100 controls…" were selected for each case identified. Why 100? What is the power of the study, for individual drugs, to detect significant ORs? [Figure 3 shows very wide confidence* *intervals for dose-response relationships for specific drugs – based on only 2 countries with this data- indicating low power]*

Reply: As some NSAIDs were infrequently used in the source population of this study, we expected a potentially low power to detect significant ORs for those individual NSAIDs more rarely used in the EU. This was indeed the case, for instance, for sulindac, which in our main analysis (based on pooled individual-level data from about 10 million NSAIDs users) was used by only 0.01% of controls. Using the approach of *Lui KJ (Am J Epidemiol 1988; 127:1064)*, we estimated that, with 100 controls per case, our main analysis including 92,163 HF cases could have only identified as significant ORs of 2.00 or more with 80% power for sulindac. The pooling of individual-level data however guaranteed us a greater power to detect significant ORs for the more frequently used NSAIDs. For instance, using the same approach as above, we estimated that our main analysis had an 80% power to detect a significant OR of 1.32 (1.17) for NSAIDs used by about 0.10% (1.00%) of controls, i.e. about as much as ketorolac (rofecoxib). Interestingly, our main analysis could have detect a significant OR associated with current use of celecoxib as low as 1.08 with an 80% power. We have added a few words in the Discussion section about this issue (page 13, lines 7-12).

Power was lower in the dose-response analysis because, as the Reviewer correctly notes, only two databases (PHARMO and THIN) were considered. This is a limitation of our study that we acknowledge in the Discussion section of the manuscript (page 15, lines 12-17).

6.5) *Cohort subjects were classified into current, recent and past NSAID use. The latter acted as the 'reference group' and included those whose prescription was more than 183 days before the index date. Might this not introduce confounding by indication? What were the possible reasons for stopping?*

Reply: We agree with the Reviewer that confounding might be a possible threat to our findings, as we acknowledge in the Discussion section. There, we state that (page 16, from line 8):

> *"Lastly, residual confounding must also be considered. This is related to the fact that some diseases that modify both the risk of HF and the probability of current NSAID use may have not been fully accounted for in this study. To protect against this possibility, all our estimates were adjusted for concomitant (i.e. in the current period) use of specific drugs (e.g. nitrates, diuretics, other drugs for CV diseases) as a proxy of patients' current health status. Still, residual confounding cannot be excluded."*

There might be several reasons for stopping NSAID treatment more than 183 days before the index date that have the potential to be associated with HF risk. One such possibility could be represented by cardiovascular diseases (CVDs) other than HF occurred after the start of NSAID therapy. This is because CVDs both i) increase the risk of HF and ii) they pose, according to current guidelines, a contraindication to NSAIDs. Nevertheless, in this study we adjusted our association estimates for use of drugs for CVDs during follow-up (i.e. in the last 90 days before the index date). This should have in part protected our findings. Additionally, any uncontrolled confounder that (like CVD) may decrease the probability of current NSAIDs use and increases the risk of HF can be expected to induce a negative association between the two and thus lead to an underestimation of the real association.

6.6) *Only covariates available in all databases were used in the statistical analysis. Does this include all features listed in Table 2? Are there any important covariates which could not be included?*

Reply: We have modified **Table 2** to indicate which covariates were available in all DBs. These always entered the conditional logistic regression models for all analyses. The remaining covariates were considered only in the database-specific analyses as now better clarified in the Methods section (page 8, starting from line 16). Additionally, we recognized that some important potential confounders we not assessed in our study. For example, in the Discussion section, we hypothesised a potential role of gout (page 16, starting from line 13) since i) gout is an independent risk factor for HF (Reference 45) and ii) NSAIDs are the first pharmacological choice for treating acute gout episodes (Reference 46). However, assuming that gout has a 1% prevalence in our source population and that it increases HF risk by 1.74-fold (References 45, 47), we estimated (Reference 49) that acute gout episodes should increase the odds of being treated with naproxen (the NSAID with the weakest statistically significant association with HF among those investigated) in the current rather than the past period by 33-fold (an implausibly high amount) to fully explain the observed naproxen-HF association. These considerations further strengthen our conclusions.

6.7) *The primary statistical analysis related to pooling individual data, whereas the secondary involved pooling summary database estimates. For the latter, results are presented for the individual drugs in the Supplementary tables. However, I could find no estimate for current use of any NSAID using the secondary approach. Was this analysis carried out? It would be useful to compare this result with the estimated OR of 1.20 obtained from the individual pooled data.*

Reply: Using the secondary meta-analytic approach we estimated that the OR for current use of any NSAIDs was 1.24 (95% CI: 1.12, 1.36), a result compatible with that obtained in the main analysis based on pooled individual-level data (OR: 1.19; 95% CI: 1.17, 1.22). We have expanded the Results section to include this finding (page 11, lines 6-7). DB-specific estimates are instead provided in **Supplementary Table S4**.

6.8) *The dose-response analysis was restricted to 2 countries only (UK and the Netherlands). Was an analysis carried out of the overall (ie. non dose-response) data for these two countries to see how it compared with the 1.20 OR for HF hospitalisation of any NSAID use for all 4 countries?*

Reply: Using pooled individual-level data from THIN (UK) and PHARMO (NL), we estimated that current users of any NSAID had a 13% higher risk of HF compared with past-users of any NSAID (OR: 1.13; 95% CI: 1.08, 1.18), a slightly lower increase in risk than that observed for all 4 countries considered together (OR: 1.19; 95% CI: 1.17, 1.22).

6.9) *A large number of individual NSAIDs are considered, and hence the problem of multiple testing and the increased possibility of chance statistically significant findings should be acknowledged.*

Reply: Following *Bender & Lange (J Clin Epidemiol 2001; 54:343)*, to address the multiple comparisons issue it is important to distinguish between the *comparisonwise error rate* (CER), i.e. the Type I error rate of each tested hypothesis considered on its own, and the *experimentwise error rate* (EER), i.e. the probability of a Type I error when all tested hypotheses are considered as components of a larger single hypothesis. Importantly, *Bender & Lange* argue that

> "*if the investigator only wants to control the CER, an adjustment for multiple tests is unnecessary.*"

As we mention in the Introduction of our paper, there is a scarcity of information on the risk of heart failure associated with the use of individual NSAIDs. Hence, in this study we assessed the association between 27 individual NSAIDs and the risk of heart failure. We assessed each of these associations as of interest in its own right rather that as part of a larger hypothesis. Hence the focus of this study was on individual associations and their quantification. Consequently we believe that, in this specific setting, controlling the CER should be of greater interest than controlling the EER. Therefore, in agreement with *Bender & Lange*, we did not perform multiple comparisons corrections in this study.

6.10) *Overall (pooled) estimates should be added to Figures 1 and 2.*

Reply: We have modified Figure 1 and 2 according to the Reviewer's suggestion.