Web appendix 1: Additional details of methods

Content

Search strategy 2
Calculation of patient-years 2
Technical implementation in WinBUGS 2
WinBUGS code for main analysis 3
Confidence levels, calculation of association between cox-2 selectivity and treatment effect, and numbers needed to treat 4
Assessment of model fit, between trial heterogeneity, and inconsistency 4
Additional analyses 5
Standard random-effects meta-analyses 6
References 6

Search strategy

We searched MEDLINE, EMBASE, and CENTRAL using truncated versions of generic and trade names of non-steroidal anti-inflammatory drugs and acetaminophen combined with the first two steps of the Cochrane methodological filter.[1] In addition, we searched proceedings of major rheumatological and oncological conferences (American College of Rheumatology, American Society of Clinical Oncology, European League Against Rheumatism, Osteoarthritis Research Society International), study registries (www.clinicaltrials.gov, www.controlled-trials.com, www.actr.org.au, www.umin.ac.jp/ctr, www.clinicaltrialresults.org), and the FDA website (www.fda.gov) using truncated versions of generic and trade names of non-steroidal anti-inflammatory drugs and acetaminophen. We manually searched reference lists of relevant articles and retrieved reports citing relevant articles via the Science Citation Index. Finally, we used the Google search engine (www.google.com) to identify additional reports and information sources by using acronyms of identified trials combined with the relevant intervention names. The search was last updated in December 2008.

Calculation of patient-years

If patient-years were not reported for a particular outcome we approximated them according to the following hierarchy:

1. Median of patient-years of other outcomes within the same trial arm
2. Patient-years = number of patients * mean follow-up duration of the trial
3. Patient-years = (number of patients not withdrawn * planned follow-up duration) + (number of patients withdrawn * planned follow-up duration / 2)
4. Number of patients * planned follow-up duration

Technical implementation in WinBUGS

The model used for all analyses and its implementation in WinBUGS is described in detail by Cooper et al.[2] It is based on multivariable Bayesian hierarchical random effects models[3] for mixed multiple treatment comparisons.[4] Basically, the model consists of two levels: the level of comparison and the level of trials. The model included a random effect at the level of trials with two basic assumptions. First, for the main analysis it was assumed that log rate ratios are from the same common distribution (see also Additional analyses below). The second assumption was that relative treatment effects add. For example, the log rate ratio comparing placebo with celecoxib was deemed to be predictable from the log rate ratios of placebo versus naproxen and naproxen versus celecoxib.

Analyses were done with Markov chain Monte Carlo simulation methods with vague prior distributions.[5] Convergence was deemed to be achieved if plots of the Gelman-Rubin statistics indicated that widths of pooled runs and individual runs stabilized around the same value and their ratio around one.[6] Given these criteria, we based our calculations on the

50,001 to 100,000 iteration, discarding the first 50,000 iterations as so-called burn-in. The median of the posterior distribution represents the rate ratio for each comparison. In analogy to 95% confidence intervals, we estimated 95% credibility intervals from the 2·5th and 97·5th percentiles of the posterior distribution. Rate ratios below one indicate a benefit of the respective preparation.

WinBUGS code for main analysis

```
model {

   for(i in 1:ns){
        for (k in (na[i]+1):3){
               dev[i,k] <- 0
        }
   }

   for(i in 1:ns){
        w[i,1] <- 0
        delta[i,t[i,1]] <- 0

# vague priors for trial baselines
        mu[i] ~ dnorm(0,0.001)

        for (k in 1:na[i]){

# likelihood function
               r[i,t[i,k]] ~ dpois(lambda[i,t[i,k]])

# evidence synthesis model
               log(lambda[i,t[i,k]]) <- log(py[i,t[i,k]]/1000) + mu[i] + delta[i,t[i,k]]

               rhat[i,t[i,k]] <- lambda[i,t[i,k]]

# deviance
         dev[i,k] <- 2*(r[i,t[i,k]]*log(r[i,t[i,k]]/rhat[i,t[i,k]]) - (r[i,t[i,k]] -
                     rhat[i,t[i,k]]))

# residuals
               res[i,t[i,k]] <- (r[i,t[i,k]] - rhat[i,t[i,k]])/sqrt(rhat[i,t[i,k]])

        }

        for (k in 2:na[i]){

# trial specific log rate ratio
               delta[i,t[i,k]] ~ dnorm(md[i,t[i,k]],taud[i,t[i,k]])

# mean log rate ratio
               md[i,t[i,k]] <- d[t[i,k]] - d[t[i,1]] + sw[i,k]

# adjustment for multi-arm trials
               w[i,k] <- delta[i,t[i,k]] - d[t[i,k]] + d[t[i,1]]
               sw[i,k] <- sum(w[i,1:k-1]) /(k-1)

# precision (inverse variance) of mean log rate ratio
               taud[i,t[i,k]] <- tau*2*(k-1)/k

        }
   }

   for (k in 1:3){

# total residual deviance
        vecresdev[k] <- sum(dev[,k])
   }
   resdev <- sum(vecresdev[])

# define placebo as reference
   d[1] <- 0

   for (k in 2:nt){
```

```
# vague priors for basic parameters
        d[k] ~ dnorm(0,0.001)
   }

# vague prior for random effects standard deviation
   sd ~ dunif(0,2)
   tau <- 1/pow(sd,2)
   tau2 <- 1/tau

}
```

Confidence levels, calculation of association between cox-2 selectivity and treatment effect, and numbers needed to treat

Posterior distributions were used for deriving confidence levels i.e. posterior probablilities, ranks and linear regression coefficients:

- A confidence level for a specific rate ratio was derived by dividing the number of iterations resulting in a rate ratio of at least the specific value divided by the overall number of iterations (50,000). For example, 6536 iterations showed a rate ratio larger than 1.2 for the outcome myocardial infarction and the comparison between naproxen and placebo. Consequently, the corresponding posterior probability is 6536 / 50,000 = 15%.
- At each iteration, the association between COX-2 specificity of each intervention[7] and the rate ratio for the corresponding intervention against placebo was derived using standard linear regression. Median values of the regression coefficient for each outcome were used to derive the final regression coefficient with 2.5% and 97.5% percentiles to derive the accompanying 95% credibility intervals.
- To calculate numbers needed to treat (NNT) we considered a population with hypothetical baseline risks.[8] The baseline risk chosen corresponds approximately to the risk in elderly patients with rheumatoid arthritis.[9,10] At each iteration, the corresponding number needed to treat was calculated using the following formula: NNT = 1 / (baseline risk – rate ratio * baseline risk). Median values of numbers needed to treat for each outcome were used to derive the final number needed to treat with 2.5% and 97.5% percentiles to derive the accompanying 95% credibility intervals.

Assessment of model fit, between trial heterogeneity, and inconsistency

We used three criteria to assess whether the model provided adequate fit to the underlying data. All are based on the residual deviance:[2,11]

- The mean of the residual deviance should be approximately similar to the number of data points used in the model. The following interpretations were used for assessment: For large numbers of data points (n) residual deviance approximately follows a chi-squared distribution with degrees of freedom given by the number of observations and with variance twice the number of data points. Mean residual deviance lying within ± 1.96 * SQRT(2 * n) of the number of data points were deemed "adequate".

- At least 95% of means of standardized node-based residuals should be within ± 1.96 of the standard normal distribution. The following interpretations were used for assessment: Model fit was deemed "adequate" if at least 95% of residuals were within ± 1.96 of the standard normal distribution.
- Normal plots of residuals lied closely around a line on visual inspection.

In our primary analysis we assumed for each outcome one common heterogeneity parameter τ^2 across comparisons. The parameter τ^2 corresponds to the variance of the underlying distribution and is difficult to interpret. However, based on a ratio of rate ratios of two randomly drawn trials from this underlying distribution, we considered $\tau^2 = 0.4$ as indicating "substantial" heterogeneity (corresponding to a ratio of rate ratios of 2.0), $\tau^2 = 0.14$ as "moderate" heterogeneity (ratio of rate ratios of 1.5), and $\tau^2 = 0.04$ as "low" (ratio of rate ratios of 1.25).[12] For each τ^2 we estimated 95% credibility intervals. To assess the robustness of results we also implemented models which assumed that τ^2 could vary across different comparisons of each active preparation with placebo (see Additional analyses below).

We used inconsistency factors as previously described to assess the consistency of the network i.e. the concordance of direct randomized comparisons within trials and indirect comparisons between trials.[11] Inconsistency factors can be interpreted as the difference between direct and indirect comparisons measured on the log rate ratio scale. Their number is restricted to the number of independent closed loops in the network. The inconsistency factors and their corresponding loops are provided in webappendix 2. For ease of interpretation we back-transformed inconsistency factors to ratios of rate ratios and expressed inconsistency as percentage difference in rate ratios between direct and indirect comparisons. Values can range from 0% to infinity. A value near 0% indicates that all the comparisons in the network are consistent, showing fully coherent estimates of rate ratios comparing any two interventions. The more the value deviates from 0%, the more inconsistent the network. Values of 100% might be interpreted as "substantial" inconsistency (corresponding to a ratio of rate ratios of 2.0), values of 50% as "moderate" (ratio of rate ratios of 1.5), and values of 25% as "low" inconsistency (ratio of rate ratios of 1.25). For each inconsistency factor we estimated 95% credibility intervals.

Additional analyses

1. Sensitivity analyses were performed by restricting the analysis to trials a) with external adjudication of events; b) in patients with musculoskeletal conditions; and c) to trials were the use of low dose Aspirin was allowed. Adequate concealment of allocation and blinding of patients and healthcare providers were not considered since all trials fulfilled these criteria. In addition, we were not able to restrict the analysis to trials with intention-to-treat analysis because only 13 trials fulfilled this criterion leaving too few trials and events for the analysis.

2. To explore the influence of our inclusion criteria we restricted the analyses 1) to trials with a minimum number of patient-years per trial arm of 500 and 2) to interventions with at least 50 accumulated myocardial infarctions in eligible trials.

3. We explored a potential dosage effect by classifying trials using dosages below the maximum approved dosage as low/moderate dose trials and trials using maximum approved dosages or higher dosages as high dose. Because the number of low/moderate dose trials was low we excluded low/moderate dose trials and restricted analyses to high dose trials.

4. A common heterogeneity parameter τ^2 was assumed across all comparisons in our main analyses and we checked the robustness of this assumption by 1) trying to implementing also a model where between-trial variance τ^2 was modeled for each comparison separately; and 2) by implementing a fixed effects model.

5. Finally, we performed analyses with trials identified as outliers excluded. Outliers were defined as trials with means of standardized node-based residuals outside ± 1.96 of the standard normal distribution.

Standard random-effects meta-analyses

We calculated Bayesian random-effects meta-analyses for all available direct comparisons.[13] Comparisons with zero events in both groups were excluded from the analysis. Rate ratios were used as measure of treatment effects.

References

1 Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; **309:** 1286-91.

2 Cooper NJ, Sutton AJ, Lu G, Khunti K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. *Arch Intern Med* 2006; **166:** 1269-75.

3 Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995; **14:** 2685-99.

4 Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004; **23:** 3105-24.

5 Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005; **331:** 897-900.

6 Brooks S, Gelman A. Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7:** 434-55.

7 Warner TD, Mitchell JA. Cyclooxygenases: new forms, new inhibitors, and lessons from the clinic. *FASEB J* 2004; **18:** 790-804.

8 Ebrahim S. Numbers needed to treat derived from meta-analyses: pitfalls and cautions. In: Egger M, Davey Smith G, Altman D, eds. Systematic reviews in health care. London: BMJ Publishing Group, 2001: 386-99.

9 Ray WA, Stein CM, Daugherty JR, Hall K, Arbogast PG, Griffin MR. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *Lancet* 2002; **360:** 1071-3.

10 Solomon DH, Goodson NJ, Katz JN, et al. Patterns of cardiovascular risk in rheumatoid arthritis. *Ann Rheum Dis* 2006; **65:** 1608-12.

11 Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006; **101:** 447-59.

12 Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health care evaluation. Chichester: John Wiley & Sons, 2004.

13 Sutton AJ. Bayesian methods in meta-analysis. In: Sutton AJ, Abrams KR, Jones DR, Jones DR, Sheldon TA, Song F, eds. Methods for meta-analysis in medical research. Chichester, UK: John Wiley & Sons, 2000: 163-90.