A Metaanalysis of Severe Upper Gastrointestinal Complications of Nonsteroidal Antiinflammatory Drugs

JOSHUA J. OFMAN, CATHERINE H. MACLEAN, WALTER L. STRAUS, SALLY C. MORTON, MARC L. BERGER, ELIZABETH A. ROTH, and PAUL SHEKELLE

ABSTRACT

Objective. Prior metaanalyses of the risk of upper gastrointestinal (GI) complications associated with nonsteroidal antiinflammatory drugs (NSAID) have focused on the published English language epidemiologic literature and/or only a portion of the relevant evidence, restrictions that are now known to be associated with bias in metaanalysis. We synthesized the published and unpublished evidence to determine the least biased estimates of the risks of perforations, ulcers, and bleeds (PUB) associated with NSAID use from all study designs and all languages.

Methods. Data sources: Using MEDLINE, EMBASE, HEALTHSTAR, and BIOSIS, we searched for English and non-English language studies of NSAID from 1966–1998 reporting primary data on GI complications. We obtained unpublished data from the US Food and Drug Administration (FDA) new drug application (NDA) reviews. NDA were hand searched to identify unpublished studies with inclusion criteria identical to those used for published reports. Study selection: Studies had to assess the use of oral NSAID for more than 4 days duration in subjects > 18 years of age and report on the clinically relevant upper GI outcomes of PUB.

Results. Two reviewers evaluated 4881 published titles and identified 13 NSAID versus placebo randomized clinical trials and 3 previously unpublished FDA placebo controlled randomized controlled trials, 9 cohort studies, and 23 case control studies sufficiently clinically homogeneous to pool. Two reviewers extracted data about study characteristics and study quality. Data synthesis: The majority of clinical trials were of good quality, but observational studies had methodologic limitations. The pooled odds ratio (OR) from 16 NSAID versus placebo clinical trials, comprising 4431 patients, was 5.36 (95% CI: 1.79, 16.1). The pooled relative risk of PUB from 9 cohort studies comprising over 750,000 person-years of exposure was 2.7 (95% CI: 2.1, 3.5). The pooled OR of PUB from 23 case control studies using age and sex matching, representing 25,732 patients, was 3.0 (95% CI: 2.5, 3.7). Data were insufficient to justify subgroup analyses stratified by age, comorbid conditions, drug, or dose.

Conclusion. These data support an association between the use of NSAID and serious upper GI complications, including estimates from different study designs. Prior pooled estimates about the effect of patient and drug variables on increased risk must be viewed with caution. (J Rheumatol 2002;29:804–12)
Our objective was to produce the least biased metaanalysis of the relationship between NSAID use and severe GI complications by assessing the effects of NSAID on PUB, including available published and unpublished data without language restrictions, and examining all study designs.

MATERIALS AND METHODS
We performed a systematic review of the medical literature for studies reporting primary data regarding the GI complications of the oral use of NSAID. We searched the MEDLINE, BIOSIS, EMBASE, and HEALTHSTAR computerized bibliographic databases from 1966 to 1998 (Table 1). Bibliographies from previous systematic reviews and personal files were also reviewed, and articles included in previous systematic reviews were screened. We searched for unpublished data by requesting data from the US Food and Drug Administration (FDA) under the Freedom of Information Act. The search was not restricted by study design or language. All stages of the review were performed independently by 2 physician reviewers trained in health services research and the principles of critical appraisal. Disagreement between reviewers was resolved by consensus.

Table 1. Search strategy.

<table>
<thead>
<tr>
<th>Major Exploded Subject Headings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Non-steroidal Anti-inflammatory Agents or NSAIDS</td>
</tr>
<tr>
<td>Or Prostaglandin Synthase Inhibitor/s And Cyclo-oxygenase Inhibitor/s</td>
</tr>
<tr>
<td>And Gastrointestinal Diseases or Esophageal Diseases or Digestive System Diseases</td>
</tr>
<tr>
<td>2. Digestive Disease/s, Digestive Complication/s, Gastrointestinal Complication/s, Dyspepsia, Peptic or Esophageal or Stomach Ulcer/s, Gastrointestinal or Digestive Hemorrhage or Bleeding or Perforation</td>
</tr>
<tr>
<td>And Adverse Toxic-Side Effect/s, Risk, Chemically Induced, Contraindications</td>
</tr>
<tr>
<td>And Human Only</td>
</tr>
<tr>
<td>Not Neoplasms, Cancer, Malignancy, Carcinoma, Adenocarcinoma/s, Case Reports, Case Studies, Case Series, Veterinary, Animal Experiment/s, Letters</td>
</tr>
</tbody>
</table>

Using a standardized form, elements of study quality were collected from unmasked articles that passed the article-screening phase, with disagreement resolved by consensus. Quality elements for all studies included comparability among groups (randomization for clinical trials), blinding, and treatment of withdrawals or missing data. Controlled clinical trials were scored for quality according to criteria developed by Jadad. These criteria were developed from published clinical trial data and relied on details of blinding and randomization, and accounting for patient dropouts. The 0 to 5 quality scoring system has been shown to discriminate between clinical trials, with a score of 3 and above indicating “high quality.” We also assessed randomized trials for concealment of allocation, but this was reported so infrequently that its discriminative value was poor. We performed duplicate abstraction with consensus resolution for study design, study population, interventions and exposures, outcome measures, features of the analysis, and study statistics.

FDA Review
FDA new drug application (NDA) reviews were obtained for the 5 NSAID representing 69.5% of the 1998 US prescription NSAID market: diclofenac, etodolac, ibuprofen, nabumetone, and naproxen (data on file). Within NDA are FDA prepared summaries of the original primary data from the studies submitted by the manufacturer in support of their application. Using the same criteria applied to published studies, we examined each NDA to identify for inclusion studies that had not been published. Data extraction was performed in duplicate with consensus resolution of discrepancies. All studies were assessed for the quality of reporting using the scale developed by Jadad.

Metaanalysis
Stratification of studies. Because the clinical trial and epidemiologic NSAID studies vary in their methods of capturing and reporting NSAID related GI events, we chose a priori to stratify the pooled analysis by study design and calculated estimates separately for each stratum.

Individual study statistics. For the randomized clinical trials and case control studies, we estimated the percentage of patients who had the outcome in the treatment and in the control groups. For cohort studies, we estimated the outcome rate per person-year observed. To assess the treatment effect, we estimated the odds ratios (OR) for randomized clinical trials (RCT) and case control studies, and the risk ratios for cohort studies.

Data considerations. One study used a crossover design and reported on a PUB outcome “6 weeks into the active drug period.” We could not determine whether this outcome occurred in the group receiving the active drug before or after the 2 week washout period. We counted this PUB as occurring during NSAID therapy. If zero outcomes were reported for a study’s control group, in order to ensure that all study statistics were defined, we applied the standard contingency table adjustment by adding a 0.5 to the treatment and control group numbers of outcomes, and a 1.0 to the numbers of patients in both the treatment and control groups.

2.7, 3.8, 4.0, and between 3.0 and 5.0 for individual drugs. While each study has strengths, these analyses have limitations including restricting the analysis to observational studies, analyzing only data available in the English language, and relying solely on the published literature. Each of these methodologic limitations has been empirically demonstrated to be associated with bias in metaanalyses.

Our objective was to produce the least biased metaanalysis of the relationship between NSAID use and severe GI complications by assessing the effects of NSAID on PUB, including available published and unpublished data without language restrictions, and examining all study designs.

MATERIALS AND METHODS
We performed a systematic review of the medical literature for studies reporting primary data regarding the GI complications of the oral use of NSAID. We searched the MEDLINE, BIOSIS, EMBASE, and HEALTHSTAR computerized bibliographic databases from 1966 to 1998 (Table 1). Bibliographies from previous systematic reviews and personal files were also reviewed, and articles included in previous systematic reviews were screened. We searched for unpublished data by requesting data from the US Food and Drug Administration (FDA) under the Freedom of Information Act. The search was not restricted by study design or language. All stages of the review were performed independently by 2 physician reviewers trained in health services research and the principles of critical appraisal. Disagreement between reviewers was resolved by consensus.

We reviewed articles for all abstracts that passed screening, as well as the articles for titles that did not have abstracts associated with them, or the article was a case report or a case series. The search strategy identified 4881 published titles (see Figure 1 for outcome of literature search). Titles were rejected only when it was clear that the article was an animal study or that it did not pertain to oral NSAID. Abstracts were obtained for all titles that passed title screening, and 2177 abstracts were reviewed. Abstracts were rejected if they met any of the following explicit criteria: failed title inclusion criteria; the article was a review, editorial, letter, clinical practice guideline, or consensus statement; or the article was a case report or a case series.

We reviewed articles for all abstracts that passed screening, as well as the articles for titles that did not have abstracts associated with them, totaling 1321 English language and 409 non-English language articles. Studies were rejected if they met any of the following explicit criteria: failed abstract inclusion criteria or the article did not report any of the following clinical GI outcomes: mortality, hospitalization for a GI event, GI bleeding, GI ulcers, and any form of GI upset (including but not limited to dyspepsia, abdominal pain, nausea, vomiting, heartburn, and pyrosis). The following report only includes results of outcomes related to PUB. We excluded studies that reported endoscopic outcomes in the absence of any relevant clinical outcomes since there are no available data to support a direct causal link between the endoscopic findings associated with NSAID use and clinical outcomes. The study population was restricted to subjects 18 years of age and older. We excluded studies in which the duration of exposure to the NSAID was less than 5 days, since such short exposures were not the focus of our clinical interest. Physicians fluent in the language of the article reviewed each non-English language article with the assistance of a study group member. We reviewed articles published in the following languages: Chinese, Danish, Dutch, Finnish, French, German, Hungarian, Italian, Japanese, Norwegian, Portuguese, Spanish, Swedish, and Russian. We did not have interpreters available for 20 studies that were written primarily in Eastern European languages. Non-English language articles did not undergo duplicate review.

Using a standardized form, elements of study quality were collected from unmasked articles that passed the article-screening phase, with disagreement resolved by consensus. Quality elements for all studies included comparability among groups (randomization for clinical trials), blinding, and treatment of withdrawals or missing data. Controlled clinical trials were scored for quality according to criteria developed by Jadad. These criteria were developed from published clinical trial data and relied on details of blinding and randomization, and accounting for patient dropouts. The 0 to 5 quality scoring system has been shown to discriminate between clinical trials, with a score of 3 and above indicating “high quality.” We also assessed randomized trials for concealment of allocation, but this was reported so infrequently that its discriminative value was poor. We performed duplicate abstraction with consensus resolution for study design, study population, interventions and exposures, outcome measures, features of the analysis, and study statistics.

FDA Review
FDA new drug application (NDA) reviews were obtained for the 5 NSAID representing 69.5% of the 1998 US prescription NSAID market: diclofenac, etodolac, ibuprofen, nabumetone, and naproxen (data on file). Within NDA are FDA prepared summaries of the original primary data from the studies submitted by the manufacturer in support of their application. Using the same criteria applied to published studies, we examined each NDA to identify for inclusion studies that had not been published. Data extraction was performed in duplicate with consensus resolution of discrepancies. All studies were assessed for the quality of reporting using the scale developed by Jadad.

Metaanalysis
Stratification of studies. Because the clinical trial and epidemiologic NSAID studies vary in their methods of capturing and reporting NSAID related GI events, we chose a priori to stratify the pooled analysis by study design and calculated estimates separately for each stratum.

Individual study statistics. For the randomized clinical trials and case control studies, we estimated the percentage of patients who had the outcome in the treatment and in the control groups. For cohort studies, we estimated the outcome rate per person-year observed. To assess the treatment effect, we estimated the odds ratios (OR) for randomized clinical trials (RCT) and case control studies, and the risk ratios for cohort studies.

Data considerations. One study used a crossover design and reported on a PUB outcome “6 weeks into the active drug period.” We could not determine whether this outcome occurred in the group receiving the active drug before or after the 2 week washout period. We counted this PUB as occurring during NSAID therapy. If zero outcomes were reported for a study’s control group, in order to ensure that all study statistics were defined, we applied the standard contingency table adjustment by adding a 0.5 to the treatment and control group numbers of outcomes, and a 1.0 to the numbers of patients in both the treatment and control groups.
For each study design, we first excluded studies that were too clinically heterogeneous to statistically pool. For example, one cohort study was excluded from statistical pooling because it evaluated the risk of PUB in only those NSAID users with pre-existing dyspepsia. Other studies were felt to be too clinically heterogeneous because they were not age or sex matched, or used outcomes such as death from PUB rather than the occurrence of PUB. We then pooled statistics across all studies and within specific strata, e.g., study design, using a random effects model, which accounts for both within and between study variance, using standard formulas for the estimates and confidence intervals. This is generally a...
more conservative method than the fixed effect model used by prior meta-
analyses, and results in wider confidence intervals. For the RCT OR, we
chose Peto’s method23,24. Along with the pooled estimates, we report the
results of the chi-square test of heterogeneity, which tests the null hypoth-
thesis that all studies are estimating the same underlying effect. Because this
test is known to have low power, nonrejection of this test does not indicate
lack of heterogeneity.

Planned subgroup analysis. We planned to assess the effect of a variety of
patient and drug variables on the association between NSAID and PUB,
including patient age, presence of prior GI problems, use of concurrent
medications such as corticosteroids, drug type, and drug dose, but would
only do so if these variables were reported in at least 70% of the studies, in
order to avoid a special form of “publication bias.”

Assessment of publication bias. We assessed the possibility of publication
bias for the RCT and cohort and case control studies. For each design separa-
tely, we evaluated a funnel plot of the log OR (RCT and case controls)
and log risk ratios (cohort studies) graphically for asymmetry resulting
from the non-publication of small, negative studies. Because graphical
evaluation can be subjective, we also conducted an adjusted rank correla-
tion test25 and a regression asymmetry test26 as formal statistical tests for
publication bias. We drew these graphs and conducted these tests in the
statistical package Stata27.

RESULTS
After excluding from this analysis articles solely or predomin-
antly about aspirin, or prophylaxis of NSAID complications
with misoprostol or other agents, we identified in the
published literature 55 NSAID versus placebo RCT28-81, 24
cohort studies8,20,82-103, and 57 case control studies3,4,104-158.
From the FDA data, we identified an additional 37 previ-
ously unpublished NSAID versus placebo clinical trials
(complete evidence tables available on request).

In general, the clinical trials were of good quality
according to the classification of Jadad, with 81% of studies
scoring 3 or greater. Most of the case control studies used
matching on age and sex (and possibly other variables) to
help enhance comparability between groups, but few studies
assessed exposure status blinded to outcomes. In contrast,
much of the cohort studies did not use matching or analytic
comparability between NSAID users and non-NSAID users,
and in 3 studies the temporal relation of NSAID exposure to
the outcome was not defined.

Summary of Effects
We estimated the risk of the more serious GI complications
of PUB due to NSAID using data from RCT and cohort
and case control studies (Table 2). Thirteen16,39,46,51,54,55,58,64-
66,70,80 published and 3 FDA NSAID versus placebo RCT
reported data on PUB. Five46,54,55,64,70 of the 13 published
trials enrolled only patients less than 65 years, while the
remainder had no upper age limit exclusion. Seven of the
clinical trials studied patients with rheumatoid arthritis, 8
trials were for 4 weeks or less in duration, and 4 trials
studied high doses of NSAID. There were no reports of PUB
in the placebo groups of these trials, whereas the percentage
of patients in the NSAID treated group was 0.2% (95% CI:
0%, 0.4%). The pooled OR is 5.36 (95% CI: 1.79, 16.1),

Table 2. Summary of results for perforations, ulcers, and bleeds.

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Number of Studies</th>
<th>Number of Patients</th>
<th>Odds (OR) or Risk Ratio (RR) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSAID vs placebo RCT</td>
<td>16</td>
<td>4431</td>
<td>OR 5.36 (1.79, 16.1)</td>
</tr>
<tr>
<td>Cohort studies‡</td>
<td>9</td>
<td>758,776</td>
<td>RR 2.7 (2.1, 3.5)</td>
</tr>
<tr>
<td>Case control studies‡</td>
<td>23</td>
<td>25,732</td>
<td>OR* 3.0 (2.5, 3.7)</td>
</tr>
</tbody>
</table>

* The total number of patients.  ‡ Chi-square test of heterogeneity, p < 0.05,
* Age and sex matched.

which is a good approximation of the relative risk since the
event rate is rare.

In the cohort data, there were 13 studies20,82,83,85,87,
89,91,92,94,97,98,102 that reported data on PUB, representing
nearly 800,000 patient-years of exposure to NSAID.
Nine20,82,83,85,87,89,92,97,98,102 of the 13 studies were judged suffi-
ciently clinically homogeneous to synthesize, and the
random effects pooled risk ratio for PUB was 2.7 (95% CI: 
2.1, 3.5). Four were excluded because the study: used death
from PUB rather than occurrence of PUB as the outcome91,
studied patients with preexisting dyspepsia rather than an
otherwise unselected population92, used a chart review to
include only those PUB occurrences the investigators felt
were “probably” caused by NSAID use94, and reported relative
risks adjusted for many variables rather than the actual
population based data99.

Among the 57 case control studies that met our criteria,
23 studies3,106-108,111-113,118,125,126,128,131,133,134,138-140,142-144,150,151,153
(contributing over 25,000 patients) were judged sufficiently
clinically homogeneous to pool (in that they all used age and
sex matching and measured PUB as an outcome). The
summary pooled OR, taken from unadjusted OR in each
study, was 3.0 (95% CI: 2.5, 3.7) (Table 2).

Subgroup Analyses
While all of the randomized clinical trials reported the drug
and dose used (but rarely age or exclusion of prior GI prob-
lems), there were too few RCT reporting PUB outcomes to
support subgroup analyses. Among the cohort and case
control studies not one variable (age, drug, dose, etc.) was
reported in more than 70% of the eligible studies. The variable
reported in the greatest proportion of studies was age in
the cohort studies. In the 9 cohort studies contributing data
to the pooled analysis, separate analyses by age were
reported for three83,97,98 and 2 studies presented analysis of the
relative risk of PUB by age, adjusted for multiple other
variables92,102. No other variable was present in even 50% of
eligible studies.

Since a recent report suggests that risk estimates may
vary depending on whether studies were performed prior to
1990159, we assessed this using our data. When stratifying
the analysis by whether studies were published before or
after 1990, there was no statistical difference in the results
stratifying the RCT or case control studies. Only the cohort data suggest that the risk estimates may be greater in the more recent literature. The risk ratio for cohort studies published before 1990 is 2.0 (95% CI: 1.7, 2.3) with 257,652 patient-years of exposure compared to studies published in 1990 or later with a risk ratio of 3.4 (95% CI: 2.8, 4.1) with 501,124 patient-years of exposure.

Assessment of Publication Bias

No evidence of publication bias was apparent for either the cohort or case control studies, either graphically and or via the adjusted rank correlation test (p = 0.99 and p = 0.92, respectively)25, or the regression asymmetry test (p = 0.41 and p = 0.45, respectively)26. Evidence of publication bias was present among the RCT both graphically and based on the formal tests (p < 0.05). The direction of the estimated bias suggested that small studies with NSAID related PUB outcomes were not reported.

**DISCUSSION**

Valid estimates of the upper GI complications of NSAID are critical for informing treatment decisions. There is a vast literature on this topic, and from 4881 titles identified in our search, our systematic review identified 16 RCT, 9 cohort studies, and 23 case control studies. This is more than previous reviews have identified. By comparison, prior studies, and 23 case control studies. This is more than previous metaanalyses have identified less than 20 total studies on risk of PUB eligible for data synthesis8,11. There is empirical evidence that study design is associated with bias, with observational studies reporting exaggerated effects compared to randomized studies14 and, within observational studies, those using a case control design reporting exaggerated effects compared to cohort studies160. Therefore, unlike previous metaanalyses, we refrained from pooling across study design. Particular strengths of this analysis include the use of published and unpublished RCT data to derive risk estimates to complement those obtained from observational studies, and the assessment of publication bias, particularly as it pertains to the degree to which data about subgroups are reported and the implications for the conclusions of prior studies reporting risk estimates in these subgroups.

The estimated risk ratio for PUB in our analysis varied from 5.4 in RCT to 2.7 in cohort studies to 3.0 in case control studies, all supporting a strong association. The width of the confidence interval in the RCT data make this source of risk estimate the least precise and least generalizable. The estimates from cohort data after 1990 are similar to the estimates from case control studies, and thus likely reflect the more accurate estimates from observational studies. Which estimate of risk is most relevant to practicing clinicians? We believe our estimate derived from cohort data is most relevant, for 2 reasons. First, while we expect that the estimate of risk would be least biased when derived from randomized placebo controlled trials, the populations being studied in such trials are usually a highly select group and are not necessarily representative of the patients seen in clinical practice. Additionally, our statistical methods indicate the likely presence of publication bias with the “missing” studies being those reporting greater risk of PUB. Second, there is empiric evidence that case control studies of diagnostic tests yield exaggerated estimates of effect compared to cohort studies160, and theoretic reasons to suspect the same will be true for observational studies of risk.

With regard to subgroup analyses, our finding that patient and drug variables such as age, drug type, dose, etc., are reported in only a small fraction of available studies is cause for some concern. The pooling of subgroups in such circumstances is likely to yield exaggerated estimates of effect, as the reporting of subgroup analyses in the original studies is almost certainly not random, but rather more likely if they are “interesting” (i.e., statistically significant)161. This phenomenon of the differential reporting of subgroup analyses or secondary outcomes is a special form of “publication bias.” Prior pooled estimates that certain patient characteristics or drug and dose regimens are associated with a lesser or greater risk of PUB must be viewed with caution. There are several limitations of this analysis. First, the validity of our results is dependent upon the quality and validity of the primary data. However, most of the clinical trials were of “good” quality according to the only validated scale of quality of controlled trials. While we did identify aspects of the observational studies that are presumed to represent methodologic weaknesses, as there are at present no criteria of the methodologic quality of observational studies that has empiric validation, we did not attempt to categorize observational studies as “high” or “low” quality, or use such a classification (other than study design) in the analysis. Also, when attempting to incorporate unpublished data, we used a sample of clinically reasonable FDA data regarding the most commonly used agents, rather than reviewing all NDA reports of all NSAID.

The strengths of this analysis include our attempts to identify all relevant studies, including unpublished data and non-English language studies, and the estimation of risk stratified by variables known to produce bias (study design). Our use of a random effects model, while not a panacea for dealing with heterogeneity162, when combined with careful assessment of clinical study differences and an a priori decision to stratify based on study design, is a more robust approach than a fixed effects pooling of all studies as previously employed. The estimates derived from this analysis are based on a greater number of studies with more patients than any prior review. We believe these are the best evidence based estimates to date regarding the risk of clinically important upper GI complications associated with NSAID. These data will be helpful to clinicians when assessing the benefits and risks of NSAID therapy for their patients and in
the evaluation of new drugs or therapies intended to reduce the risk of NSAID associated upper GI complications.

REFERENCES


19. Data on file. IMS America, Westport, CT, USA.


Lussier A, Tetreault L, Lebel E. Comparative study of


