Characterization of gastrointestinal adverse effects reported in clinical studies of corticosteroid therapy

Tone Westergren[a,*], Sigrid Narum[b], Marianne Klemp[c]

[a]Deptartment of Pharmacology, Regional Medicines Information & Pharmacovigilance Centre (RELIS), Oslo University Hospital Rikshospitalet, PO Box 4950 Nydalen, 0424 Oslo, Norway
[b]Centre for Psychopharmacology, Diakonhjemmet Hospital, PO Box 23 Vinderen, 0319 Oslo, Norway
[c]Department of Pharmacology, University of Oslo, PO Box 1057 Blindern, Oslo 0316, Norway

Abstract

Objectives: To examine whether 159 studies included in a previous meta-analysis reported on gastrointestinal bleeding or perforation in accordance with the CONSORT extension for reporting harms outcomes (CONSORT Harms recommendations checklist); whether differences were associated with funding source, journal, or publication year; and whether the CONSORT Harms checklist is a suitable tool for evaluation of adverse effects reporting.

Study Design and Setting: Articles were assessed for fulfillment of the CONSORT Harms recommendations, funding source, publication type, and year. Agreement between reviewers was assessed by comparing scores for each study.

Results: The mean CONSORT Harms score was 5.25 out of 10 (standard deviation $\pm$ 2.09). Most studies included information on participant withdrawals (133 studies, 83.6%), absolute risk of gastrointestinal bleeding or perforation (130 studies, 81.8%), and how harms-related information was collected (118 studies, 74.2%). Reporting of gastrointestinal bleeding or perforation increased with higher scores (odds ratio 1.173, $P = 0.042$). There was no significant association between CONSORT Harms score achieved and publication year or funding source, but there was a trend toward higher scores in studies published in the major medical journals (score difference 0.78, $P = 0.052$). Definitions of gastrointestinal bleeding differed between studies. Reviewer agreement was fair to moderate with large variations.

Conclusion: Few studies in the systematic review received high scores using the CONSORT Harms criteria. Most studies reported on the most important criteria regarding risk of gastrointestinal bleeding or perforation. Reviewer agreement showed large variations due to imprecise texts and ambiguous criteria. Routine scoring according to fulfillment of the CONSORT Harms recommendations would be inadvisable without qualified judgment. © 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Gastrointestinal hemorrhage; Glucocorticoids; Pharmacovigilance; Adverse drug reaction reporting systems/standards; Guideline adherence; Systematic review

1. Introduction

Most randomized clinical trials are designed to evaluate efficacy of drug treatment and therefore provide better assessments of benefits than risks. However, comprehensive and reliable data on both benefits and risks are necessary to make a balanced risk/benefit assessment. Safety and risk of adverse effects cannot be thoroughly explored in short-term studies that include only a limited patient group. Shortcomings in adverse effects monitoring and reporting may lead to inadequate assessments and lower estimates of serious harm [1,2]. The Declaration of Helsinki [3], developed by the World Medical Association, states that medical research may only be conducted if the importance of the objective outweighs the risk to the research subjects. Failure to identify relevant risks may lead to research projects with an unacceptable risk/benefit balance. If problems of unsystematic monitoring or reporting of adverse effects are added to inconsistent or heterogeneous data, it may be

What is new?

Key findings

- Studies included in a previous review on the risk of gastrointestinal bleeding or perforation during corticosteroid treatment were analyzed with regard to quality of adverse effects monitoring and reporting. The studies were assessed and scored using the CONSORT Harms criteria with 10 recommendations.

- The mean score was 5.3/10, which means that several CONSORT Harms criteria were not met for many of the studies.

- Only 59/159 studies were identified as having addressed and monitored gastrointestinal adverse effects judging from the study descriptions. However, the absolute risk of gastrointestinal bleeding or perforation was found in 130/159 studies. Gastrointestinal adverse effects were reported in studies that did not specify the intention to address them.

- Reporting of gastrointestinal bleeding or perforation was higher in studies with higher CONSORT Harms criteria scores, compared to studies with lower scores. Exclusion of studies with low scores would have led to exclusion of relevant findings of cases of gastrointestinal bleeding or perforation.

What this adds to what was known?

- This is the first in-depth analysis of adverse effects monitoring and reporting in studies that were included in a systematic review of risk of adverse effects. Data on adverse effects could be found in most studies, although several aspects of adverse effects reporting were heterogeneous and unsystematic with regard to definitions, method of monitoring, and data analysis. The study provides an insight into the realities of summarizing literature on adverse effects.

What is the implication and what should change now?

- Use of checklists is advocated for quality assessment of included clinical trials in reviews and meta-analyses. Routine scoring of clinical studies using CONSORT Harms criteria for harms assessment would be inadvisable without adding qualified judgment on the study in question. Published clinical studies generally do not fulfill all criteria in the CONSORT Harms checklist. Too narrow inclusion criteria may eliminate studies that are suboptimal with regard to adverse effects reporting but still give relevant information on adverse effects. Conclusions about adverse effects made in systematic reviews should take the variability and heterogeneous reporting of adverse effects in the underlying data into account.

impossible to draw conclusions regarding risk from single or pooled clinical studies and to perform systematic reviews for risk/benefit assessment [4]. Weaknesses in the original reporting of adverse effects will be magnified when those reports form the basis of meta-analyses and systematic reviews.

The Cochrane Handbook for Systematic Reviews of Interventions emphasizes the need for careful scrutiny of the studies' intensity of monitoring adverse effects and clarity of reporting [5]. The PRISMA statement addresses improvements in quality and transparency of systematic reviews by way of minimum standards for reporting [6], and a PRISMA Harms extension for systematic reviews has been developed [7]. Use of the GRADE approach for grading quality of evidence [8] is recommended by the *British Medical Journal* and the Cochrane Collaboration among others but does not provide the tools for a detailed examination of the adverse effects reporting. Other methods have been proposed to address the quality of adverse effects or harms reporting. Both the CONSORT group and Cochrane Adverse Effects Methods group advocate the use of checklists when including clinical studies for methodology review or meta-analysis [5,9]. These are lists of recommendations describing what information should be included in various parts of the article. A commonly cited example is the CONSORT checklist [10], an initiative to improve the reporting of clinical trials, with an added 10 recommendations for reporting harms published in 2004 [11], often referred to as the CONSORT Harms criteria. Others have developed extended, more detailed versions [12]. The McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm) covers many of the same recommendations as the CONSORT checklist [13]. As yet, there are no universally endorsed instruments for assessing risk of bias with regard to adverse effects or harms in clinical trials or systematic reviews.

We have previously published a systematic review and meta-analysis of corticosteroid use and risk of gastrointestinal bleeding or perforation [14], including only randomized, double-blinded studies. During the review process, it became clear that the included 159 studies varied widely in their descriptions and methods of adverse effects reporting and definitions of gastrointestinal bleeding, although they all fulfilled our inclusion criteria. We have analyzed the studies to examine whether they reported on adverse effects in accordance with the CONSORT extension for reporting harms outcomes (referred to as CONSORT Harms criteria) [11]; to examine whether any differences could be linked to variables such as funding source, journal quality, or publication year; and to evaluate whether the CONSORT Harms criteria are a suitable tool for evaluation of the quality of adverse effects reporting in clinical trials.

2. Methods

2.1. Study data and criteria assessments

One hundred fifty-nine articles included in a previous systematic review and meta-analysis of corticosteroids and risk of gastrointestinal bleeding or perforation (referred to as gastrointestinal bleeding in the rest of the article) were included in the analysis [14]. A standardized checklist and data extraction form was prepared based on the CONSORT Harms recommendations. The criteria were discussed by all authors to arrive at a common understanding.

We collected data on 10 different outcomes using the CONSORT Harms recommendations (Table 1, recommendations 1–10). Two recommendations (3 and 8) were modified to include gastrointestinal adverse events only, to reflect whether the study specified assessments of adverse gastrointestinal effects associated with study treatment, as this was the adverse effect addressed in the meta-analysis [14]. The relevant text from the articles was extracted and scored as 0 or 1 by two of the authors independently (S.N. and T.W.) by interpreting the checklist criteria in relation to the article text. Several CONSORT Harms criteria included two or more parameters. If the article met any one of the criteria that CONSORT included for a topic, it was counted as fulfilled for that topic, as has been practiced elsewhere [15]. The scores were discussed and a final score was decided. The reviewers were not blinded to the name of the journal or the authors.

All articles were assessed for reporting of gastrointestinal adverse effects, funding source, publication type, and year. Studies scoring 8, 9, or 10 were classified as high-score studies. Studies scoring 3 or less were classified as low-score studies. To see if publication of the CONSORT Harms extension in 2004 had led to improved adverse effects reporting, studies were grouped according to publication year (≤2004, ≥2005). Studies with industry coauthorship or donations of product or money were classified as industry sponsored. Studies published by one of the five major medical journals (*Lancet*, *British Medical Journal*, *New England Journal of Medicine*, *Journal of the American Medical Association*, and *Annals of Internal Medicine*) were analyzed separately.

2.2. Reviewer agreement

Agreement between reviewers was used as an indicator of the ease of use and suitability of the CONSORT Harms recommendations. Interrater agreement for each study was analyzed using Gwet's agreement coefficient with first-order chance correction, AC1 (value 0-1) [16]. Interrater agreement for each CONSORT Harms criterion across the 159 studies was analyzed using Gwet's AC1 [17].

2.3. Statistical analysis

We calculated correlations using the Pearson chi-square test and differences in scores using the *t*-test for equality of means. Logistic regression analysis was used to examine the relationship between CONSORT Harms criteria scores and the likelihood of reporting gastrointestinal bleeding. Correlations, score comparisons, and logistic regressions were analyzed using IBM SPSS Statistics (version 23). All analyses were two tailed, with an α of 0.05.

3. Results

3.1. Study scores using CONSORT Harms criteria

The 159 clinical studies each received a total score for 10 different criteria, giving a total of 3,180 criteria assessments in the two separate reviewer evaluations and 1,590 criteria assessments evaluated for the final score. All discrepancies were resolved during the final discussion, and no cases were referred to the third author.

In the final assessment, the studies received a mean score of 5.25 out of a maximum of 10 (standard deviation [SD] $\pm$ 2.09). For studies without a subgroup analysis (excluding recommendation 9), the mean score was 5.15 (SD $\pm$ 1.97) out of a maximum of 9. Most studies did not include a subgroup analysis.

The distribution of criteria scores among the studies is shown in Fig. 1. Logistic regression analysis showed a higher reporting of gastrointestinal bleeding with increasing CONSORT Harms criteria scores (odds ratio [OR] 1.17, 95% confidence interval 1.01–1.37, $P = 0.042$). The odds of reporting cases of gastrointestinal bleeding were three times higher for high-score studies compared to low-score studies (OR 3.43, 95% confidence interval 1.17–10.04).

The recommendations with the highest scores were recommendation 6—participant withdrawals (133 studies, 83.6%), 8—absolute risk of gastrointestinal adverse events (130 studies, 81.8%), and 4—clarify how harms-related information was collected (118 studies, 74.2%). The recommendations with the lowest scores were recommendation 9—subgroup analysis (16 studies, 10.1%), 2—collection of harms data mentioned in introduction (48 studies, 30.2%), and 5—plan for presenting and analyzing information on harms (51 studies, 32.1%). The scores according to the CONSORT Harms recommendations are presented in Table 2.

Fifty-nine studies (37.1%) did address and monitor for gastrointestinal adverse events, either specifically or as part of a comprehensive clinical examination (recommendation 3). The remaining 100 studies (62.9%) did not address gastrointestinal adverse effects or did not describe a clinical examination of sufficient extent. Despite this, the absolute risk of gastrointestinal adverse events (recommendation 8) could be found in 130/159 studies (81.8%). This number includes studies with zero observed gastrointestinal adverse

Table 1. Scoring criteria

Recommendation 1. If the study collected data on harms and benefits, the title or abstract should so state
Definition: Score 1 if any mention of harms, adverse events, side effects, toxicity, or complications, excluding those clearly due to lack of treatment effects or underlying disease. If not, score 0.
Recommendation 2. If the study collected data on harms and benefits, the introduction should so state
Definition: Score 1 if any mention of harms, adverse events, side effects, toxicity, or complications, excluding those clearly due to lack of treatment effects or underlying disease. If not, score 0.
Recommendation 3. List addressed gastrointestinal adverse events with definitions for each
Definition: Score 1 if any gastrointestinal adverse event was specified as an outcome to be addressed or if the clinical examination described is perceived as comprehensive enough to discover overt gastrointestinal adverse effects and any other major events. If not, score 0.
Recommendation 4. Clarify how harms-related information was collected
Definition: Score 1 if method of collection or system of monitoring for harms is specified. If not, score 0.
Recommendation 5. Describe plan for presenting and analyzing information on harms
Definition: Score 1 if harms analysis is specified, or if the general method of result analysis appeared to have been applied to harms data. If not, score 0.
Recommendation 6. Describe for each arm the participant withdrawals that are due to harm and the experience with the allocated treatments
Definition: Score 1 if withdrawals due to adverse events were specified. If not, score 0.
Recommendation 7. Provide the denominators for analyses on harm
Definition: Score 1 if denominators are described. If not, score 0.
Recommendation 8. Present the absolute risk of each gastrointestinal adverse event and present appropriate metrics for recurrent events, continuous variables, and scale variable
Definition: Score 1 if absolute risk can be found for any gastrointestinal adverse effect. If not, score 0.
Recommendation 9. Describe any subgroup analyses and explanatory analyses for harms
Definition: Score 1 if any subgroup analysis for adverse drug reactions was done. If not, score 0.
Recommendation 10. Provide a balanced discussion of benefits and harms with emphasis on study limitations, generalizability and other sources of information on harms
Definition: Score 1 if the discussion is perceived as balanced and study limitations are discussed. If not, score 0.

Adapted from CONSORT Harms recommendations [11].

effects, which in several cases had to be interpreted from lists of observed adverse effects or statements of no detected adverse effects. In 29 studies (18.2%), the number of patients included in the risk analysis was not described. However, cases of gastrointestinal bleeding were reported in five of those publications. In studies where gastrointestinal bleeding was addressed or observed, the definitions and descriptions varied widely. A detailed description is provided in Supplementary Materials.

Twenty-four studies (15.1%) received a score of 8, 9, or 10 and were classified as high-score studies. Those studies included 4,510 patients (2,277 receiving steroid, 2,233 receiving placebo), of which 16 studies (66.7%) reported cases of gastrointestinal bleeding (155 cases in the steroid group, 92 cases in the placebo group). Twelve of the 24

Fig. 1. Studies grouped by CONSORT Harms criteria scores (*N* = 159).

high-score studies (50%) concerned prevention of bronchopulmonary dysplasia in pediatric patients and contributed a major proportion of cases of gastrointestinal bleeding (120 cases in 1,066 corticosteroid-treated patients, 69 cases in 1,047 placebo-treated patients).

Thirty-eight studies (23.9%) received a score of 3 or less, indicating that few of the CONSORT Harms criteria were met in the publications (low-score studies). Those studies included 6,605 patients (3,312 receiving steroid, 3,293 receiving placebo), of which 14 studies (36.8%) reported cases of gastrointestinal bleeding (41 in the steroid group, 22 in the placebo group). Twenty-four studies (63.2%) did not report any cases of bleeding. The main reasons for achieving low scores were that adverse effects were not mentioned in title, abstract, or introduction; plans for presenting and analyzing harms were not described; or discussions were not perceived as balanced. None of the low-score studies received a score for addressing gastrointestinal adverse effects. Still, in most of the low-score studies, it was possible to present an absolute risk of gastrointestinal bleeding (25/38, 65.8%) and participant withdrawals due to harm (23/38, 60.5%).

3.2. CONSORT Harms criteria scores relating to key variables

We found no significant correlation between the CONSORT Harms score and publication year, ambulant or hospitalized patients, or funding source (industry sponsored or not) (Table 3). There was a trend toward higher scores for
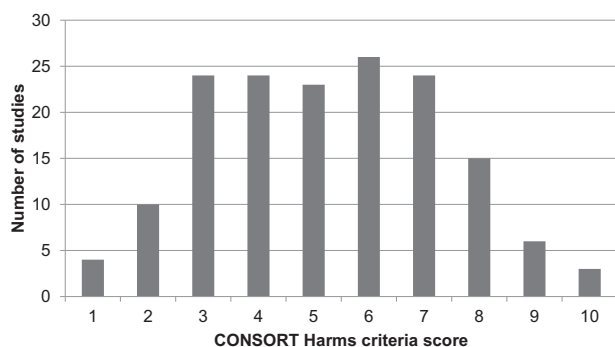
Table 2. Studies which fulfilled CONSORT Harms criteria, *N* = 159

CONSORT Harms criterion	No. (%)
1. If the study collected data on harms and benefits, the title or abstract should so state	93 (58.5)
2. If the study collected data on harms and benefits, the introduction should so state	48 (30.2)
3. List addressed adverse gastrointestinal events with definitions for each	59 (37.1)
4. Clarify how harms-related information was collected	118 (74.2)
5. Describe plan for presenting and analyzing information on harms	51 (32.1)
6. Describe for each arm the participant withdrawals that are due to harm and the experience with the allocated treatments	133 (83.6)
7. Provide the denominators for analyses on harms	102 (64.2)
8. Present the absolute risk of each gastrointestinal adverse event and present appropriate metrics for recurrent events, continuous variables, and scale variable	130 (81.8)
9. Describe any subgroup analyses and explanatory analyses for harms	16 (10.1)
10. Provide a balanced discussion of benefits and harms with emphasis on study limitations, generalizability, and other sources of information on harms	85 (53.5)

studies published in the major medical journals, with mean score 5.86 vs. 5.08 in other journals (*P* = 0.052). The studies with the highest scores (score ≥ 8) had 33.3% (8/24) industry sponsoring, compared to 54.8% (74/135) industry sponsoring for the rest of the studies (*P* = 0.052).

To see if reporting had improved in the most recent years, the reporting after 2007 was analyzed separately. Studies published in 2007–30.6.2011 (*N* = 26) had a mean score of 4.88. Studies published in the major medical journals in 2007–30.6.2011 (*N* = 7) had a mean score of 5.29.

3.3. Qualitative assessment

Several studies collected data on adverse effects, including gastrointestinal, without mentioning the fact in title, abstract, or introduction. In many studies, adverse effect monitoring had obviously been performed without mention of intention or method. Risk of gastrointestinal adverse effects had often been considered beforehand, as evidenced by exclusion criteria such as previous peptic ulceration, but not mentioned in methods, results, or discussion sections. Plans for presenting and analyzing information on harms were often not specified. Information on adverse effects was in many cases presented less systematically than efficacy outcomes and could be found in various sections of the publications. In some studies, efficacy and harm were analyzed in the same way; in other studies, statistical methods were applied to efficacy outcome only. Some studies limited adverse effects reporting to the most common or most serious cases. Denominators were sometimes specified for efficacy only, not for adverse effects, and could only be found by inference by comparing adverse effects tables with text. Several studies presented adverse effect data as percentages. If the denominator for adverse effect analysis was not clearly stated, the absolute risk could not be found. In several studies that quantified withdrawals, the reason was not always stated but could be inferred by interpreting the text in relation to the withdrawal data. Conclusions of safety, such as "no safety problems," were sometimes drawn despite underpowered study design and unsystematic addressing of adverse effects.

3.4. Reviewer agreement

In the analysis by two separate reviewers, the mean CONSORT Harms criteria scores were 5.19 (SD ± 2.13) and 6.06 (SD ± 2.11), respectively, for the 159 studies. Interrater agreement for each study, calculated as Gwet's AC1, had a mean value of 0.56 (SD ± 0.29) and a median value of 0.62 (range −0.28 to 1.00). The 15 studies with slight or poor reviewer agreement coefficients (Gwet's AC1 < 0.2) received significantly lower CONSORT Harms scores than studies with higher degrees of agreement (3.87 vs. 5.40) (*P* = 0.007). Interrater agreement for each CONSORT Harms criterion through all 159 studies, using Gwet's AC1 agreement coefficient, showed a mean value of 0.58 (SD ± 0.15) and a median value of 0.57 (range 0.32-0.82). Agreements between reviewers differed with regard to individual criteria. The criteria with the three lowest Gwet's AC1 scores were recommendations 7, 9, and 10 (0.42, 0.32, and 0.46, respectively).

Details can be found in Supplementary Materials.

4. Discussion

4.1. CONSORT harms criteria score, main findings

We examined the reporting of gastrointestinal adverse effects in 159 published randomized controlled trials which were included in a published meta-analysis addressing risk of gastrointestinal bleeding associated with corticosteroid use [14]. The studies had undergone quality assessment and fulfilled the criteria for inclusion in a systematic review. However, analysis of the publications, using criteria proposed in CONSORT Harms adjusted for gastrointestinal adverse effects, showed that few studies received high scores. Adverse effects monitoring and reporting varied greatly, and most of the studies did not fulfill several criteria. Only 24 studies (15.1%) received a score of 8 or more. The mean CONSORT Harms criteria score of 5.3 for all 10 criteria corresponds generally to that found by Maggi et al. [18]. Exclusion of criterion 9 had

Table 3. Correlation between CONSORT Harms scores and key variables

Key variable	No (%)	Mean CONSORT Harms score	Score difference (95% confidence interval)	P
Industry sponsored	82/159 (51.6)	5.00	0.52 (−0.13 to 1.17)	0.118
Not industry sponsored	77/159 (48.4)	5.52		
Score ≥ 8, industry sponsored	8/24 (33.3)	8.25	0.38 (−0.27 to 1.02)	0.239
Score ≥ 8, not industry sponsored	16/24 (66.7)	8.63		
Published in major medical journal	35/159 (22.0)	5.86	0.78 (−0.01 to 1.56)	0.052
Published in other journals	124/159 (78.0)	5.08		
Published 1983−2004	114/159 (71.7)	5.26	0.04 (−0.69 to 0.77)	0.912
Published 2005−30.6.2011	45/159 (28.3)	5.22		
Ambulant patients	56/159 (35.2)	5.32	0.11 (−0.58 to 0.80)	0.757
Hospitalized patients	103/159 (64.8)	5.21		

only a limited effect on the overall mean score, reflecting that relatively few studies had received a score on this criterion.

Some of the criteria were fulfilled for most of the studies but, in many cases, to a limited degree where information had to be inferred by the reviewers. The present study gives no indications as to why criteria were not fulfilled. Most of the 159 studies focused on treatment efficacy. Adverse effects were generally given little space and were, for most studies, not a prespecified end point. Another possible explanation may be journal text limitations, although space limitations should not be an excuse to exclude information on this highly important issue when reporting on results of a clinical study. It remains to be seen whether adverse effects reporting will improve with increasing use of electronic publications.

4.2. Weaknesses in monitoring and reporting gastrointestinal bleeding

Occurrence of gastrointestinal bleeding was assumed to be an objective and unambiguous adverse effect that would have been described if observed in the studies. Most studies did not address the risk of gastrointestinal bleeding specifically, although several studies did record gastrointestinal bleeding and discussed the risk in the introduction section. Definitions of gastrointestinal bleeding varied widely and cases could possibly be hidden within broader diagnostic groups such as "gastrointestinal reactions." This may be an even greater problem with more subjective adverse effects.

Some of the CONSORT Harms criteria may be less critical than others when it comes to the facts of whether the study did address gastrointestinal adverse effects and whether any adverse effects were reported. Many studies did monitor adverse effects, including gastrointestinal, with little mention of intention or method. This was a major reason for interrater differences on recommendation 4. It can be argued that the most important recommendations regarding actual findings of gastrointestinal bleeding risk are recommendations 6−8 (withdrawals, denominators, and absolute risks), although it is reasonable to expect

any intention to look for adverse effects to be mentioned in the abstract or introduction. Information on absolute risks was given in 130 studies (81.8%), although not always clearly stated. In addition, some studies reported cases of gastrointestinal bleeding without describing absolute risk. One hundred thirty-three studies (83.6%) described withdrawals and experience with the allocated treatments to some extent. Recommendations 1−5 (stating of intention and plans for analyzing harms data) were not always fulfilled, even when adverse effects were described in the results sections. Subgroup analysis (recommendation 9) is obviously not a quality criterion for reporting harms if not part of the study. Several of the studies that received low scores using the CONSORT Harms criteria nevertheless gave an impression of thoroughness and awareness of the risk of adverse effects, despite the fact that little space was allotted to adverse effect descriptions in the publication.

Superficial descriptions of adverse effects and use of cutoff valuations such as "serious" or "frequency > 5%" make it possible to avoid describing all adverse effects that occurred. It has been argued that it is safer to assume that adverse effects were not ascertained or not recorded than to assume that the prevalence or incidence was zero if the adverse effect is not mentioned specifically [19]. However, in a clinical trial, there are risks of several adverse effects and it would be unreasonable to expect authors to mention all those that did not occur, unless they were addressed specifically.

Studies with low quality of reporting of harms, as assessed using the CONSORT Harms criteria, might have a correspondingly lower chance of finding adverse effects, from either poor study design or poor monitoring.

Inclusion of only those studies that described active or comprehensive adverse effects monitoring would have eliminated 100 of 159 studies and 63 cases of gastrointestinal bleeding from our systematic review [14]. If mentioning of adverse effects in title, abstract, or introduction sections had been a selection criterion, 57 studies would have been lost for analysis. Exclusion of studies with low scores would have led to exclusion of relevant findings of cases of gastrointestinal bleeding.

4.3. CONSORT Harms scores in relation to key variables

We found no clear correlation between publication year (before or after publication of the CONSORT Harms criteria) and the reporting of adverse effects. This reflects most previous findings [20–23], whereas Haidich et al. [24] found a somewhat increased reporting of harms from 2003 to 2006. In our analysis, studies published in 2007–30.6.2011 had a lower mean score than studies published in the period preceding publication of the CONSORT Harms criteria, indicating that the reporting of adverse effects did not improve over time.

Contrary to expectations, there was a relatively small score difference between studies published in major medical journals and other journals. There was a trend toward a higher mean score for these studies. Haidich et al. [24] analyzed randomized clinical trials published in the five major medical journals and found mean scores of 5.8 and 6.7 for studies published in 2003 and 2006, respectively. This corresponds generally with the mean score of 5.86 found in our study, but in our study, the scores appeared to decline over time. In an analysis of studies published in four major medical journals in 2009, Maggi et al. [18] found that most studies did not incorporate the CONSORT Harms recommendations sufficiently.

In contrast to previous studies, where industry-funded studies have shown better safety reporting than nonindustry studies [18,20,24,25], we found a trend toward worse safety reporting in studies that were supported or funded by the pharmaceutical industry. This may be due to our broad definition of sponsoring or the fact that most studies were published before 2004 and were probably not performed for regulatory purposes, as the corticosteroid used had been on the market for several years.

4.4. Reviewer agreement

Analysis of initial reviewer agreement for each study and for each CONSORT Harms criterion across studies showed fair-to-moderate agreement with large variations. Low agreement was mainly caused by differences in interpretation of information in the article texts and difficulties in determining whether a criterion was sufficiently fulfilled or not. In addition, many of the CONSORT Harms criteria include several questions within one recommendation. Some authors have addressed the ambiguity by splitting some of the original recommendations into several, more precise subcategories [20,22,26,27], in some cases with option of half credits [24,27]. Because of the heterogeneity of the studies regarding the methods descriptions and the presentation of data, a more detailed approach using a more specific checklist would probably not have reduced the necessity for judgment or resulted in greater agreement between reviewers. The use of half credits if a criterion was partly fulfilled might have resulted in more specific scores, but there would still be an element of judgment regarding the degree of fulfillment of each criterion.

Because subgroup analysis of adverse effects is rarely done, other authors have excluded CONSORT Harms recommendation 9 from assessment [20,26,27]. Subgroup analysis of harm was done in several of the studies included in our review but with focus on harm as a result of disease or treatment failure. This was a major reason for score discrepancies between reviewers.

4.5. Limitations

Several studies reported adverse effects without mentioning gastrointestinal bleeding. As all the studies did address or report adverse effects to some extent, we concluded that no gastrointestinal bleeding occurred in those studies. This assumption may be mistaken, as a lack of reports does not necessarily mean that the adverse effects did not occur [9]. In studies where gastrointestinal bleeding was not observed, the nonoccurrence cannot necessarily be expected to be commented on unless the adverse effect was expected or looked for. There is, however, an uncertainty if the risk profile is not described in detail. Our finding of higher reporting of gastrointestinal bleeding with increasing CONSORT Harms criteria scores might indicate underreporting of adverse events in the low-score studies.

We scored the studies through assessment by at least two authors. However, application of the CONSORT Harms criteria to clinical studies involves considerable judgment. Other reviewers may differ in their opinion as to what should constitute a score of 0 or 1.

The recommendations of CONSORT [10] and CONSORT Harms [11] were developed to improve the quality of clinical study reporting and were not intended as a validated tool for assessing the methodological quality of studies. A validated tool is not available at the present time.

5. Conclusion

Analysis of clinical studies included in a previous review and meta-analysis, using criteria proposed in CONSORT Harms adjusted for gastrointestinal adverse effects, showed that few studies received high scores. Reporting of gastrointestinal bleeding increased with increasing CONSORT Harms score. Application of the CONSORT Harms criteria to the clinical studies involved considerable judgment, because of the multiple items within several of the criteria and the highly variable adverse effects reporting in the studies. So far, no clear assessment method has been proposed to describe studies adequately without risking eliminating studies with relevant findings. In our opinion, routine scoring by CONSORT Harms criteria for harms assessment would be inadvisable without adding qualified judgment on the study in question.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.10.018.

References

[1] Ross SD. Drug-related adverse events: a readers' guide to assessing literature reviews and meta-analyses. Arch Intern Med 2001;161: 1041—6.

[2] Yazici Y. Some concerns about adverse event reporting in randomized clinical trials. Bull NYU Hosp Jt Dis 2008;66(2):143—5.

[3] The World Medical Association (WMA). WMA Declaration of Helsinki—Ethical principles for medical research involving human subjects. Available at https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/. Accessed November 21, 2017.

[4] Saltman D, Jackson D, Newton PJ, Davidson PM. In pursuit of certainty: can the systematic review process deliver? BMC Med Inform Decis Mak 2013;13:25.

[5] Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (Updated March 2011). The Cochrane Collaboration; 2011. Available at http://handbook.cochrane.org/Chapter 14: Adverse effects. Accessed November 12, 2014.

[6] PRISMA transparent reporting of systematic reviews and meta-analyses. PRISMA Checklist. 2009. Available at http://www.prisma-statement.org/. Accessed November 12, 2014.

[7] Zorzela L, Loke YK, Ioannidis JP, Golder S, Santaguida L, Altman D, et al. PRISMA Harms: improving harms reporting in systematic reviews. BMJ 2016;352:i157.

[8] Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

[9] Loke YK, Price D, Herxheimer A. Systematic reviews of adverse effects: framework for a structured approach. BMC Med Res Methodol 2007;7:32.

[10] CONSORT Transparent reporting of trials. CONSORT checklist. Available at http://www.consort-statement.org. Accessed November 19, 2013.

[11] Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004;141:781—8.

[12] Pitrou I, Boutron I, Ahmad N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 2009;169:1756—61.

[13] Agency for Healthcare Research and Quality (AHRQ). Methods guide for effectiveness and comparative effectiveness reviews. AHRQ Publication No 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2014. Available at https://www.ncbi.nlm.nih.gov/books/NBK47095/. Accessed November 12, 2014.

[14] Narum S, Westergren T, Klemp M. Corticosteroids and risk of gastrointestinal bleeding: a systematic review and meta-analysis. BMJ Open 2014;4(5):e004587.

[15] Lee PE, Fischer HD, Rochon PA, Gill SS, Herrmann N, Bell CM, et al. Published randomized controlled trials of drug therapy for dementia often lack complete data on harm. J Clin Epidemiol 2008;61: 1152—60.

[16] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(Pt 1): 29—48.

[17] Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013;13:61.

[18] Maggi CB, Griebeler IH, Dal Pizzol Tda S. Information on adverse events in randomised clinical trials assessing drug interventions published in four medical journals with high impact factors. Int J Risk Saf Med 2014;26(1):9—22.

[19] Chou R, Aronson N, Atkins D, Ismaila AS, Santaguida P, Smith DH, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010;63:502—12.

[20] Shukralla AA, Tudur-Smith C, Powell GA, Williamson PR, Marson AG. Reporting of adverse events in randomised controlled trials of antiepileptic drugs using the CONSORT criteria for reporting harms. Epilepsy Res 2011;97(1—2):20—9.

[21] Nuovo J, Sather C. Reporting adverse events in randomized controlled trials. Pharmacoepidemiol Drug Saf 2007;16(3):349—51.

[22] Carmichael K, Nolan SJ, Weston J, Tudur Smith C, Marson AG. Assessment of the quality of harms reporting in non-randomised studies and randomised controlled studies of topiramate for the treatment of epilepsy using CONSORT criteria. Epilepsy Res 2015;114: 106—13.

[23] Meister R, von Wolff A, Mohr H, Nestoriuc Y, Härter M, Hölzel L, et al. Adverse event methods were heterogeneous and insufficiently reported in randomized trials on persistent depressive disorder. J Clin Epidemiol 2016;71:97—108.

[24] Haidich AB, Birtsou C, Dardavessis T, Tirodimos I, Arvanitidou M. The quality of safety reporting in trials is still suboptimal: survey of major general medical journals. J Clin Epidemiol 2011;64:124—35.

[25] Smith SM, Chang RD, Pereira A, Shah N, Gilron I, Katz NP, et al. Adherence to CONSORT harms-reporting recommendations in publications of recent analgesic clinical trials: an ACTTION systematic review. Pain 2012;153(12):2415—21.

[26] Bagul NB, Kirkham JJ. The reporting of harms in randomized controlled trials of hypertension using the CONSORT criteria for harm reporting. Clin Exp Hypertens 2012;34:548—54.

[27] Breau RH, Gaboury I, Scales CD Jr, Fesperman SF, Watterson JD, Dahm P. Reporting of harm in randomized controlled trials published in the urological literature. J Urol 2010;183:1693—7.