Role of systematic reviews in the identification and prevention of poor quality research and scientific misconduct: random reflections based on anaesthesiologic cases

ELIA, Nadia

Abstract
Research in clinical medicine often suffers from poor methodological quality and is sometimes based on fraudulent data, which may bias the conclusions of systematic reviews. This dissertation develops the idea that not all research malpractices and misconduct threaten the scientific knowledge, nor may they all bias the conclusions of systematic reviews to the same extent, but rather that the strict procedures applied during proper systematic reviewing actually should make systematic reviews robust to some common errors and misconduct. Moreover, well-performed systematic reviews could become part of the solution to the problem by providing clear research agendas and by providing important methodological information that may be used for future research. Furthermore, systematic reviewers may correct the published literature after having identified errors during critical appraisal of articles, and they may even play the role of whistle-blowers by identifying new, yet unrecognised, cases of misconducts.

Reference
ELIA, Nadia. Role of systematic reviews in the identification and prevention of poor quality research and scientific misconduct: random reflections based on anaesthesiologic cases. Thèse de privat-docent : Univ. Genève, 2017

DOI : 10.13097/archive-ouverte/unige:92532

Available at:
http://archive-ouverte.unige.ch/unige:92532

Disclaimer: layout of this document may differ from the published version.
"Role of systematic reviews in the identification and prevention of poor quality research and scientific misconduct: random reflections based on anaesthesiologic cases"

Thesis submitted to the Faculty of Medicine of the University of Geneva

for the degree of Privat-Docent

by

Nadia ELIA

Geneva

2016
“There is something extraordinarily satisfying in designing an RCT of “place of therapy”, writing the protocol in such a way as to avoid all the ethical pitfalls, persuading all the necessary people to participate, and checking to see that no one cheats”.

Archibald Cochrane, 1972
Summary

Research in clinical medicine often suffers from poor methodological quality and is sometimes based on fraudulent data. Although systematic reviews are generally recognised as generating the highest level of evidence of the impact of an intervention through the synthesis of all existing research on a topic, a broad area of research has developed to highlight that some common malpractices and major misconduct in original reports may bias the conclusions of systematic reviews, questioning their reliability in this context.

In this dissertation, I develop the idea that not all research malpractices and misconduct threaten the scientific knowledge, nor may they all bias the conclusions of systematic reviews to the same extent, but rather that the strict procedures applied during proper systematic reviewing actually should make systematic reviews robust to some common errors and misconduct. Moreover, I raise the idea that well-performed systematic reviews could become part of the solution to the broad problem of poor quality research and research misconduct by providing clear research agendas which would highlight what research is, respectively is not, needed, and thereby decreasing the number of redundant trials performed and published and by providing important methodological information that may be used for future research. Furthermore, systematic reviewers may play an important role by correcting the published literature after having identified errors during critical appraisal of articles, and they may even play the role of whistle-blowers by identifying new, yet unrecognised, cases of misconducts. Finally, I suggest that performing of a systematic review during the academic curriculum could improve the methodological research understanding of young physician-researcher.

Although most of the answers to the difficult problem of research malpractice and misconduct probably lie within academic structures, policies and regulations, systematic reviews may be seen as an interesting tool to support better quality in the published medical literature.
# Table of Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td>9</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>11</td>
</tr>
<tr>
<td>1.1 Context</td>
<td>11</td>
</tr>
<tr>
<td>1.2 Aims and structure of the dissertation</td>
<td>11</td>
</tr>
<tr>
<td>2. The experimental clinical trial</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Steps from the idea to publication</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Potential malpractices</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Safeguards</td>
<td>14</td>
</tr>
<tr>
<td>3. Honest error or misconduct?</td>
<td>19</td>
</tr>
<tr>
<td>4. The systematic review</td>
<td>21</td>
</tr>
<tr>
<td>4.1 Malpractices that do not threaten the conclusions of systematic reviews</td>
<td>21</td>
</tr>
<tr>
<td>4.2 Malpractices that do threaten the conclusions of systematic reviews</td>
<td>23</td>
</tr>
<tr>
<td>4.3 Additional roles</td>
<td>23</td>
</tr>
<tr>
<td>5. Personal contributions in this area of research</td>
<td>27</td>
</tr>
<tr>
<td>5.1 Ketamine for post-operative pain - a quantitative systematic review of randomised trials</td>
<td>29</td>
</tr>
<tr>
<td>5.2 Fate of articles that warranted retraction due to ethical concerns: a descriptive cross-sectional study</td>
<td>41</td>
</tr>
<tr>
<td>5.3 Susceptibility to fraud in systematic reviews: lessons from the Reuben case</td>
<td>51</td>
</tr>
<tr>
<td>5.4 How do authors of systematic reviews deal with research misconduct in original studies? A cross-sectional analysis of systematic reviews and survey of their authors</td>
<td>63</td>
</tr>
<tr>
<td>5.5 Ability of a meta-analysis to prevent redundant research: systematic review of studies on pain from propofol injection</td>
<td>75</td>
</tr>
<tr>
<td>6. Discussion</td>
<td>89</td>
</tr>
<tr>
<td>6.1 What should be done?</td>
<td>89</td>
</tr>
<tr>
<td>6.2 Where do systematic reviews fit into the picture?</td>
<td>92</td>
</tr>
<tr>
<td>7. Conclusion</td>
<td>94</td>
</tr>
<tr>
<td>8. Bibliography</td>
<td>95</td>
</tr>
</tbody>
</table>
Foreword

One morning in early 2009, as I arrived in my office at the Department of anaesthesiology of the University Hospitals of Geneva, I found my boss and mentor wearing a strange look on his face. We had a problem. Dr Scott S Reuben, an eminent American anaesthesiologist, had admitted having falsified and fabricated data in 21 of his articles, published between 1996 and 2008. These articles were going to be retracted. This was like an earthquake in the area of anaesthesiology. The news was immediately, and noisily, disseminated. Scott Reuben had published numerous original research articles on the topic of postoperative pain management, and anaesthesiologists were questioning what should still be regarded as “known” in this field. More worrying to me was the fact that an eminent anaesthesiologist, P.F. White, suggested that “the retraction of Dr. Reuben’s articles would compromise every meta-analysis, editorial and systematic review of analgesia trials that included these fabricated findings”. I felt particularly concerned since I had been one of the author and co-author of such systematic reviews and meta-analysis. The following years confirmed that the “Reuben Case” was unfortunately not isolated, with the emergence of two new scandals in anaesthesiology: Joachim Boldt in 2010, and Yoshitaka Fuji in 2011. One positive effect of these scandals was to motivate me, with my background in anaesthesiology, epidemiology and public health, to investigate these issues further. What was this all about? Who were these doctors that “cheated” in their research publications? Was there a way to identify fraudulent data? Also, I started to wonder whether P.F. White’s concerns could be justified: where should systematic reviews and meta-analyses stand in the context of fraudulent science? These questions have modelled my research interest ever since, and will be the object of the present thesis.
1. Introduction

1.1 Context

Experimental research in clinical medicine often suffers from poor methodological quality,\textsuperscript{10,11,12,13,14} and is sometimes based on fraudulent practices.\textsuperscript{15,16,17,18} These errors and misconduct (further called “malpractices”) threaten the quality of patient care,\textsuperscript{19,20,21,22,23} damage the reputation of science,\textsuperscript{24,25} and may also threaten the conclusions of systematic reviews relying on these trials.\textsuperscript{8,7,26} Although systematic reviews are generally regarded as generating the highest level of evidence of the impact of an intervention through the synthesis of all existing research on a topic, a broad area of research has developed to highlight the impact that some common malpractices in original reports may have on the conclusions of systematic reviews, questioning their reliability in this context.\textsuperscript{27,28,29,30,31,32} Less attention has been paid to the fact that not all research malpractices may bias the conclusions of systematic reviews, that systematic reviews may sometimes correct some of these malpractices, and that systematic reviewers may even act as whistle-blowers in the context of research misconduct.

1.2 Aims and structure of the dissertation

The aim of this dissertation is to discuss the impact of the different research malpractices on the conclusions of systematic reviews and explore what additional roles systematic reviewers could play in this context. I will start by describing the steps underlying a clinical trial and illustrate the potential malpractices as well as the safeguards that exist throughout the research process. For clarity purposes, I will focus on the experimental clinical research design. Secondly, I will shortly discuss the fuzzy limit between “honest error” and “major misconduct”, and highlight that reprehensible misconduct does not always threaten the scientific knowledge or the conclusions of systematic reviews. Thirdly, I will briefly describe the steps underlying a systematic review to illustrate why some malpractices in the included trials remain harmless to the conclusions of the systematic review, how systematic reviewers may identify and sometimes correct these malpractices, and how systematic reviewers may even identify new, previously unrecognised, cases of intentional misconduct. The fourth part will report my personal contributions to this area of research and I will end by discussing the future roles of systematic reviewers in the context of research malpractice.

Systematic reviews are often criticized or misunderstood.\textsuperscript{33} Although problems related to malpractices and misconduct when performing a systematic review do exist, and are becoming more frequent,\textsuperscript{34} I will not developed these in this dissertation, as it would represent a topic on its own. On the contrary, my aim here will be to highlight some of their strengths that are often overlooked or ignored. Therefore the assumption underlying the entire dissertation is that the systematic review is well performed, and sticks to the standards of high quality methodology.
2. The experimental clinical trial

In an experimental trial, as opposed to an observational study, the investigator administers an intervention to a group of patients not previously exposed to the intervention. In a randomized controlled trial (RCT), some patients receive the intervention, while others do not, and this is determined by chance. The strength of RCTs is to distribute all potential confounding variables, both known and unknown, equally among the groups, so that the difference in outcome between the groups can be attributed to the intervention only. RCTs represent the gold standard design for assessment of the impact of an intervention. Correctly performing a RCT requires adherence to a number of rules and procedures.

2.1 Steps from the idea to publication

A research starts with an idea which is translated into a good research question, that needs to be both relevant and answerable.55,36 This is followed by the elaboration of a research protocol describing the progress of the research in detail. The protocol is an official document that guides the researcher throughout the trial, can be amended, and must be submitted to different regulatory bodies for approval. The protocol should justify the choice of the study design, give a clear description of the population investigated, of the intervention administrated, the comparator chosen, and of the primary and secondary outcomes of interest, to minimise the risk of subsequent biases. The number of patients required to reach a conclusion also needs to be justified based on a pre-hoc sample size calculation. This is especially important in experimental trials applying an intervention of unknown impact to participants as it would be unethical to include either too few,37,38 but also too many patients in such a trial. The research protocol requires to be submitted to an ethics committee for formal approval in order to guarantee the safety of included patients, to check that patients are properly informed of the aims of the trial, of their rights to participate (or not) and also of their rights to change their mind regarding their participation, any time, without justification. The ethics committee should also verify the relevance, feasibility and adequacy of the methods proposed. If the trial involves the administration of a new medication, or of an old medication for a new indication, the protocol should also be submitted to the national agency for drug regulation (in Switzerland: Swissmedic). If external financing of the trial is needed, the protocol may also be submitted to a sponsor (sometimes private actors). All instances review the protocol separately, and all approvals are required before patient enrolment can start.

After formal approval of the protocol by all regulatory bodies and co-authors, the protocol needs to be published or registered in an open-access database such as, for example, “ClinicalTrial.org”. Patients fulfilling inclusion criteria can then be approached, and those who agree to participate are asked to sign an informed consent form before they can be enrolled in the study and randomized. Any deviation from the initial protocol needs to be documented; amendments need to be submitted for approval to the regulatory bodies, and all documents regarding these changes
should be completely and transparently reported in a “Trial Master File”. Once data collection is over, the results need to be analysed, interpreted and any potential biases discussed. Finally, the results of the trial are clearly and transparently reported in a scientific article, and submitted for publication to a journal. Journal editors send the submitted article to experts (called: peer-reviewers) for their opinion; peer-reviewers may suggest to reject the article or to propose ways to improve it. In case of rejection, the article can be improved and submitted to another journal for consideration. Eventually, the article gets published and will either remain un-read or un-cited or will be read, cited, commented and, in rare cases, retracted. The major steps underlying a randomized controlled trial are illustrated in Figure 1.A.

2.2 Potential malpractices
Malpractices, ranging from harmless innocent errors to intentional misconduct, may occur throughout the process of the clinical trial. The first possible problem may be to start a research based on a research question that is irrelevant. Although science requires that experiments be repeated to demonstrate the consistency of results over time and place, a good research question should aim to provide an answer that is not yet known and requires a careful review of the published literature on the topic; this is still too often lacking. Performing a trial when the answer to the research question is already known raises ethical problems, especially in the field of experimental research on humans. The limit between a relevant and a redundant research question however, is less clear. It has become highly improbable that a clinical trial could start without a research protocol at all, since the protocol needs to be submitted to regulatory bodies. However the written protocol may be incomplete or remain unregistered and inaccessible. Also, the choice of an inadequate study design and the lack, or unclear definition, of primary and secondary outcomes are frequently described. Studies comparing published research with their protocol showed that over 60% had at least one primary outcome that had changed, was introduced or omitted compared to the protocol, even in high impact factor journals. Also the use of non-validated scales for the measurement of outcomes, and the lack of pre-hoc sample size calculation are common problems. It has been shown that less than 5% of the studies in anaesthesiology reported a pre-hoc sample size calculation before 1985; that proportion has grown to about 55% in 2001 and, in a review of RCTs published in the European Journal of Anaesthesiology, we showed that it had further improved to 74% 10 years later. Improvements in the quality of published RCTs have been described by others. It is surprising however that a quarter of the RCTs published in anaesthesiology still do not justify their sample size. During the data collection stage, malpractices can take various forms, which can be regrouped under the headings of “information biases”, when the variables or outcomes are systematically recorded differently across the groups due to, for example, lack or violation of blinding, or
“selection biases” when, for example in an RCT, concealment of the randomisation code is violated. Finally, unreported deviations from the protocol have also been described.\textsuperscript{46,54} During the data analysis stage, authors may be tempted to change the primary outcome in favour of one that is statistically significant,\textsuperscript{47} to apply an inadequate statistical test to show more impressive results, or to selectively report some results\textsuperscript{55,56,57,58} or subgroup analyses,\textsuperscript{59} just to name a few very common malpractices.\textsuperscript{60} Over-interpretation of the study results, or making conclusion that are not supported by the data have also been reported frequently.\textsuperscript{61} Any \textit{intentional} attempt to hide or improve the appearance of the results, for example by ignoring outliers, is known as “data falsification”. In some rare cases, an article may be published based on inexistent and invented data; this is known as data fabrication.\textsuperscript{16,17}

Many problems may occur when writing the article, such as, authorship issues (guest- or ghost-authorship),\textsuperscript{62} not-reporting an author’s or a sponsor-related conflict of interest transparently,\textsuperscript{19,63,64} but also plagiarism of sentences from previously published work. Self-citation of one’s own work or selective citations of articles supporting one’s own view\textsuperscript{65,66} or of statistically significant results,\textsuperscript{67} is almost the rule, and has been suggested to be a gender issue.\textsuperscript{59} Simultaneous submission of the same article to different journals, which may lead to multiple publications of the same data, still happens, although prohibited. Finally, maybe the worst case for an article, is to end-up not being published at all;\textsuperscript{69,70} this has been shown to be frequent even for RCTs, in particular when RCTs show non-significant results.\textsuperscript{71,72} Recent initiatives, such as the mandatory registration of research protocols, or the “Restoring Invisible and Abandoned Trials (RIAT)” initiative, have emerged to counter-balance this problem.\textsuperscript{73} Potential malpractices during the process of a RCT are summarized in Table 1.

\subsection*{2.3 Safeguards}

Safeguards to prevent or correct malpractices during the research process exist. They include controls made to the protocol (before the trial starts), audits and monitoring performed during data collection, peer-reviews of the article (before publication), comments by cautious readers (after publication), and eventually publication of corrections, or retraction of the published article. Furthermore, the research team and academic rules and attitude towards research integrity play an important role that may influence the risk of malpractices.

Protocol submission to regulatory bodies usually represents a step where problems are highlighted, and corrected, thanks to the reviews performed by the ethics committee, the drug regulation agency or funders. This remains a difficult task and a large variety of standards have been described across ethics committees.\textsuperscript{74} Attempts to harmonise procedures are emerging;\textsuperscript{75,76} however, the impact of ethical peer-review on subsequent trial quality remains unclear.\textsuperscript{41,77,78} Sometimes, authors \textit{fail to seek} any formal ethical approval or patient consent; this may hide even more serious forms of malpractices.
Box 1
The Reuben case: uncovering fraud through the lack of ethics committee approval.
Dr Scott S Reuben was an American anaesthesiologist with an international reputation built on his research focused on multimodal analgesia for the prevention and control of post-operative pain. He was the chief of the acute pain clinic in Baystate Medical Center, Springfield, Massachusetts, when a routine review of research abstracts discovered, in May 2008, that two of the studies that he intended to present during an internal « research week » had not received formal approval by an ethics committee.2 An investigation was launched by his institution and, a year later, Dr Reuben admitted having fabricated data, and invented patients, in 21 research articles published during a 13 years period. Dr Reuben assumed the entire responsibility for this fraud, and his co-authors were not further bothered. He was finally sentenced to six months in jail and three years of supervised release. He was also ordered to repay about $400’000 after he pleaded “guilty to health care fraud”, according to Patrick Johnson from The Republican (June 2010).

The prospective registration of research protocols has become mandatory for publication in most peer-reviewed journals. Although clinical trial registries do not yet control the quality of the research protocol, this procedure should force authors to stick to their initial ideas, and prevent them from switching endpoints or introducing new endpoints that were accidentally found to be more interesting, because statistically significant. 43,79,80 It should also help to reduce non-publication of “negative” trials, and facilitate the work of peer-reviewers.81 Some journals encourage and accept to publish research protocols. In this case, the protocols benefit from a peer-review, and may be improved.82 Audits can be performed by the ethics committee or by the drug regulation agency any time during data collection. These however, are not systematic, are performed on a random basis, but have been shown to sometimes identify and correct some malpractices.83,4 Monitoring of the trial can be performed by an independent structure. This consists of a regular data quality control and is performed on demand by the trial investigators.
As illustrated in Table 1, peer-reviewers are given a central role to guarantee the good quality of the published literature. However, it has been shown that they remain unable to detect most major errors in RCTs.84,85,86 Although RCTs published in high impact factor journals have been shown to be more likely to report on methods to prevent biases compared to those published in lower impact factor journals, which could be attributed to better peer-review, this does not guarantee low risk of bias.87 More worrying is the fact that training of peer-reviewers may not be of any use,85,88,89 and worse is the recent discovery of peer-review fraudulent organisations, where some authors manage to review their own papers through the use of faked email addresses.90,91,92 Softwares are now used by most journals to detect plagiarism,93 and the occurrence of copy-paste like types of plagiarism is likely to decrease. Plagiarism of ideas though
will remain a problem. Sometimes, errors or malpractices are discovered by cautious readers after publication, eventually leading to correction, or retraction of the article.

**Box 2**

**The Boldt case: uncovering fraud by cautious readers.**

Prof Joachim Boldt was a renowned German anaesthesiologist. He had built his reputation through his publications focusing on the use of colloids. He was the head of the department of anaesthesiology at the Klinikum Ludwigshafen, a hospital in Germany. In December 2009, the journal *Anesthesia & Analgesia* published an article authored by Prof Boldt, which was followed by three letters from readers to the Editor in Chief of the journal. These letters all highlighted that the data reported in this paper were rather unlikely (too small standard deviations for the IL-6 concentrations). The Editor in Chief of the journal, who agreed with the reader’s concerns, identified the responsible entity for ethical conduct of research in Klinikum Ludwigshafen (LAK-RLP) who started an investigation on that trial. About a year later, after having identified numerous errors in the article, the article was retracted. The investigation concluded that the study had been fabricated; none of the patients or analyses had been identified. Prof Boldt admitted having fabricated the data and having signed for his co-authors on the copyright transfer forms. Following on that case of fraud, Prof Boldt's institution started an internal investigation on his work over the past 10 years; their conclusion lead to the retraction of about 90 articles published in 18 journals.

The proportion of retracted articles has increased rapidly, and a large proportion were shown to be due to misconduct, highlighting both low barriers to publication of fraudulent articles, and easier retraction practices. A "retraction index" showed that the chance of an article being retracted was higher when published in a high impact factor journal. However, it is widely believed that retracted articles only represent the tip of the iceberg, with most of the malpractices remaining un-identified, and even some identified frauds remaining un-retracted. This problem will be discussed further in chapter 5.3. Table 1 summarizes the potential malpractices and safeguards discussed above.

* computed as the number of retractions multiplied by 1,000, divided by the number of published articles within these journals during the same time period
Figure 1
Steps underlying a randomized controlled trial and a systematic review

### A. RANDOMIZED CONTROLLED TRIAL

- **Idea**
- **Research question**
  - Study design
  - Population (inclusion/exclusion)
  - Intervention / comparator
  - Primary / secondary outcomes
  - Sample size, plan of analysis

**Protocol**

- Patient enrollement (Information, consent)
- Randomisation, concealment
- Data collection

**Data collection**

- Data analyses
- Data interpretation, conclusions
- Writing and publishing the article

### B. SYSTEMATIC REVIEW

- **Idea**
- **Research question**
  - Plan of literature search
  - Inclusion / exclusion criteria
  - Intervention / comparator
  - Primary / secondary outcomes
  - Plan of analysis

**Protocol**

- Article retrieval
- Critical appraisal
- Double data extraction
- Contact with study authors

**Data collection**

- Data syntheses +/- meta-analyses
- Data interpretation, conclusions, agenda
- Writing and publishing the article

**Ethics Committee**

**Drug Regulation Agency**

**Sponsor**

**Registration**

**Possible audits and monitoring**

**Peer-reviewers**

**Editors**

**Readers**
<table>
<thead>
<tr>
<th>Steps in a RCT</th>
<th>Potential malpractices</th>
<th>Institutional safeguards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining the research question</td>
<td>Clinically irrelevant question</td>
<td>Protocol peer-review by: Ethics committee, National agency for drug regulation, Funding agencies</td>
</tr>
<tr>
<td></td>
<td>Lack of a clear research question</td>
<td>Protocol publication in journal</td>
</tr>
<tr>
<td></td>
<td>Incomplete review of existing literature</td>
<td></td>
</tr>
<tr>
<td>Writing the research protocol</td>
<td>Wrong design</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ill-defined primary and secondary outcome</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Use of non-validated outcome measures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lack of sample size calculation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lack of analysis plan</td>
<td></td>
</tr>
<tr>
<td>Submitting the research protocol</td>
<td>Failure to submit the protocol for approval</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Failure to seek patient informed consent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Failure to register protocol</td>
<td></td>
</tr>
<tr>
<td>Performing the research</td>
<td>Selection bias</td>
<td>Audits, Monitoring</td>
</tr>
<tr>
<td></td>
<td>Information bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deviation from protocol</td>
<td></td>
</tr>
<tr>
<td>Analysing the results</td>
<td>Changing the primary outcome</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong statistical tests applied</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data falsification (i.e. ignoring outliers)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data fabrication</td>
<td></td>
</tr>
<tr>
<td>Interpreting the results</td>
<td>Biased and/or over-interpretation of results</td>
<td>Peer-reviewers, Editors</td>
</tr>
<tr>
<td>Writing the article</td>
<td>Not reporting protocol violation</td>
<td>Plagiarism software</td>
</tr>
<tr>
<td></td>
<td>Selective reporting of outcomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ghost / guest authorship</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not declaring conflicts of interest</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Plagiarism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Selective or self-citation</td>
<td></td>
</tr>
<tr>
<td>Publishing the article</td>
<td>Simultaneous submission to multiple journals</td>
<td>Editors, Peer-reviewers</td>
</tr>
<tr>
<td></td>
<td>Duplicate publication</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Failure to publish research</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Falsifying the peer-review process</td>
<td></td>
</tr>
<tr>
<td>Post-publication</td>
<td>Non-correction of identified errors</td>
<td>Editors and publishers</td>
</tr>
<tr>
<td></td>
<td>Non-retraction of articles that warrant retraction</td>
<td></td>
</tr>
</tbody>
</table>
3. Honest error or misconduct?

It is usually admitted that minor errors ought to be identified and corrected, but authors don’t need to be blamed or punished. However, some errors are considered so important that they cannot be corrected, and because they are misleading, or considered misconduct, retraction of the article becomes necessary.

The limit between “honest error” and “misconduct” relies on the intention of the authors: a misconduct being an intentional error, as apposed to honest error supposedly “unconscious”.

Therefore, differentiating errors (who deserve correction) from misconduct (who deserve correction and sanction) becomes a subjective and moral matter, leading to various and inconsistent grading of the different malpractices. Initially, misconduct used to describe only data fabrication, falsification, and plagiarism, which were undoubtedly intentional malpractices. Some authors have thereafter introduced the concept of “questionable research practices” to describe more common (maybe less conscious) malpractices, and some have further tried rating misconduct from “minor” to “major” misconduct. (Fig 2)

Classification of different malpractices as either minor errors or major misconduct remains difficult. For example, the lack of a clear definition of the primary outcome can be an unconscious error made by an inexperienced researcher, but can also be an intentional malpractice made in the hope to manipulate the data during data analysis. The same can be said of most other malpractices discussed above. A less moral, and perhaps more objective way to consider the problem would be to look at it from the point of view of “Science”. In this case, we realise that first, honest errors can falsify the scientific knowledge just as much as some major misconduct can, since both may lead to conclusions that do not reflect reality, and secondly that not all recognised misconduct threaten the scientific knowledge. For example, plagiarism of someone’s wording or ideas without acknowledgement, or guest authorship, although unfair to the original authors, illegal and reprehensible, may not affect the validity of the scientific knowledge. This is good news for systematic reviewers, since not all research malpractices will bias their conclusions.

In the next chapter, I will describe the process of a systematic review and discuss whether each of the above cited malpractices could bias the conclusions of systematic reviews, and describe how the process of systematic reviewing may correct or identify new misconducts.
Figure 2
Varying definitions of research misconduct

What is Research misconduct?
*Richard Smith*

A preliminary taxonomy of research misconduct.

**Serious research misconduct**
- Fabrication: invention of data or cases.
- Falsification: wilful distortion of data.
- Plagiarism: copying of ideas, data, or words without attribution.
- Failing to get consent from an ethics committee for research.
- Not admitting that some data are missing.
- Ignoring outliers without declaring it.
- Not including data on side effects in a clinical trial.
- Concluding research in humans without informed consent or without justifying why consent was not obtained to an ethics committee.
- Publication of post hoc analyses without declaration that they were post hoc.
- Gift authorship.
- Not attributing other authors.
- Redundant publication.
- Not disclosing a conflict of interest.
- Not attempting to publish completed research.
- Failure to do an adequate search of existing research before beginning new research.

**Minor research misconduct**

**Scientific integrity, misconduct in science**
*Emilio Bossi*

Table 1
Some categories and examples of scientific misconduct.

- Misconduct that distorts scientific knowledge,
  as a result of which society can be put at risk:
  falsification (incl. deliberate withholding of data);
  fabrication of nonexistent data.

- Misconduct which misleads the scientific community:
  plagiarism;
  unjustified authorship;
  duplicate publication;
  deliberate false evaluation of projects and results.

- Questionable research practices,
  which cast doubt on the seriousness of the research:
  sloppy handling of data;
  division of the results, with publication in several different journals solely for the purpose of increasing the list of publications.

Swiss Medical Weekly 2010; 140(13-14):183-186

Research misconduct - The grey area of Questionable Research Practices
*René Custers*

4. The systematic review

The structure of a systematic review is very similar to that of a clinical trial. (Fig1.B) The process starts with an idea, and the formulation of a clear research question, and is followed by the elaboration of a research protocol describing the major steps of the review. The first of these steps is a transparent and exhaustive search of the literature for all published and unpublished articles potentially answering the research question. These articles are selected according to pre-defined inclusion criteria, and all potentially relevant articles are critically appraised, regarding validity and research methodology. Outcomes of interest are extracted by at least two authors separately to control for potential data extraction errors and contact with study authors is often necessary to clarify the methodology used, or to unearth hidden (unreported) outcomes. The results are then synthetized at the study level based on the summary measures reported in the article, and when deemed adequate, a weighted pooled summary estimate of results can be computed, known as a “meta-analysis”. If such pooling of the data does not seem reasonable, the systematic review will remain descriptive, also called “qualitative systematic review”. Interpretation of the review’s findings is then made, together with a thorough discussion of the strengths and limitations of the analyses. The review ends with a conclusion, sometimes with recommendations, but most of all, with a research agenda highlighting what is known, and what needs further clarification. The process leading to the publication of the finalised article is very similar to that of a clinical trial.

Since systematic reviews rely on previously performed trials, their conclusions obviously can suffer from some of the above-discussed malpractices. Interestingly, some of these malpractices remain harmless to their conclusions, either because they don’t directly threaten the scientific knowledge, or because the strict procedures applied during systematic reviewing have the potential to detect and to correct some of the malpractices, or take them into account during the analyses. Additionally, systematic reviewers may play an important role by suspecting or identifying misconduct in original trials, but also to improve the design and conduct of subsequent research.

4.1 Malpractices that do not threaten the conclusions of systematic reviews

Some malpractices do not threaten the scientific knowledge, and therefore do not threaten the conclusions of a well-performed systematic review either. Examples of such malpractices include guest-authorships, or plagiarism of ideas or wording of others. Although reprehensible, these malpractices are unlikely to bias the conclusions of systematic reviews. Even failure to obtain formal ethical approval for the conduct of a clinical trial, or failure to obtain patient’s informed consent, if the trial has been otherwise well performed, should not bias the conclusions of a systematic review.
Some other malpractices may theoretically threaten the scientific knowledge, at the article level, but may not be such a threat to the conclusions of systematic reviews either because they are not considered by systematic reviewers, or because they may be detected, and actions may be taken to avoid making biased conclusions. For example, the bad formulation of a research question is unlikely to be a problem since a research based on an “unanswerable” question is likely to remain un-published, one based on a “clinically irrelevant” question is likely not to be selected during the review process, and redundant researches will tend to increase the power of a meta-analysis, and the generalizability of the conclusions of a systematic review.

Also, good systematic reviewing implies the critical appraisal of retrieved articles, and trials using an inadequate study design will most likely not be retained for analyses. Changing the primary outcome during data analysis will also remain mostly harmless because systematic reviewers include trials based on predefined methodological criteria, rarely on the basis of their primary outcomes (although this can happen) and never on the statistical significance of the reported results. Systematic reviewers may detect changes made to the primary outcome by looking at the registered protocol, if available, and perform sensitivity analyses accordingly. Since quantitative systematic reviews recalculate all measures of impact of the tested intervention using the reported summary estimates at the study level, the choice of an inadequate statistical test in the primary report is unlikely to bias the results of a quantitative systematic review. Although underpowered trials are likely to report an overestimated intervention effect, their potential impact when included in a systematic review remains a matter of debate. Interestingly, 70% of the systematic reviews published in the Cochrane database were shown to be based only on underpowered studies. Moreover, systematic reviews may be viewed as designs that can make sense of underpowered trials, therefore justifying their publication since they may be subsequently included in a larger meta-analysis, and allow the estimation of a plausible effect that may be challenged in future studies. Systematic biases during patient enrolment and/or data collection can certainly bias the estimate of the effect of an intervention, and it is the difficult duty of systematic reviewers to critically appraise all retrieved trials and to decide whether or not biases may have distorted the reported results and therefore justify the non-inclusion of the trials into the analyses. Some trials fulfilling all inclusion criteria may be finally excluded if a major bias has been identified. For minor risks of biases, that are not judged severe enough to justify the exclusion of the article from the review, systematic reviewers are encouraged to rate their risk of biases, using different scores (for example: Jadad scale, modified oxford scale, or Cochrane risk of bias tool to name a few). This quality rating allows the performance of subgroup analyses according to different risk of biases, and helps quantify the impact of the bias of the review’s conclusion. The choice of the scale used however remains unclear to many authors. Personal contact with study authors is a very important part of a systematic review as it allows clarification of methodological issues, of ill-defined primary and secondary outcomes and also to unearth unreported outcomes. It has been shown to be a valuable way to decrease
selective outcome reporting. Other typical examples of research malpractices that remain mostly harmless to the conclusions of systematic reviews include citation malpractices (selective or self-citation) and over-interpretation of results by study authors, or conclusions that are not based on the reported results. The latest are harmless to quantitative systematic reviews since the interpretation of the study authors are barely read by systematic reviewers. However, qualitative systematic reviews may be more sensible to over-interpretation of results; this will be illustrated in chapter 5.4.

Non-publication of completed research could bias the scientific knowledge and the conclusions of systematic reviews towards an over-estimation of the impact of an intervention, but also sometimes towards an underestimation of this effect and there are clinically relevant examples in the medical literature of non-publication of studies with statistically significant adverse effects. Systematic reviewers are encouraged to either search for unpublished trials, and/or to test for “publication bias” through visual examination of funnel plots in an attempt to correct these malpractices. Whether statistical correction of an assumed publication bias is a good idea remains a matter of debate. Finally, duplicate publication, meaning the publication of the results of the same trial in more than one article, can be a threat to the scientific knowledge and to systematic reviews if it remains un-recognised. Systematic reviewers have been shown to be well trained in detecting such duplicate reports and therefore are able to identify them, exclude the duplicate and report on the redundant publication.

4.2 Malpractices that do threaten the conclusions of systematic reviews

Some malpractices however may bias the scientific knowledge and the conclusions of systematic reviews since they are likely to remain unnoticed by systematic reviewers. These include data fabrication and falsification, but also non-declaration of protocol violation (when the protocol is not prospectively registered or accessible), of author’s conflict of interest, ghost authorships (if associated with conflict of interest), corrupted peer-review, or failure to correct identified errors or to retract articles that ought to be retracted. This list only considers the most common malpractices. Systematic reviewers should therefore consider these potentially misleading malpractices when discussing the results of their review.

Interestingly, recent cases of massive frauds in anaesthesiology have highlighted that some cases of malpractices and misconduct may be inter-related. For example, the lack of ethical approval may hide other science threatening misconduct, such as data fabrication, as was illustrated by the cases described in Box 1 and 2. Therefore, even the above-mentioned malpractices do not directly impact of the scientific knowledge, may potentially be symptoms of more science threatening practices.

4.3 Additional roles

The previous paragraphs have illustrated why systematic reviews are vulnerable to research
malpractices and misconduct, but probably less so than original researches. Interestingly, systematic reviewers may play additional roles in this context. First, systematic review can be used as a design to *investigate* on the prevalence of research misconduct in the published literature,\(^1\) or the impact of different malpractices on the trial’s or on the systematic review’s conclusions\(^2\) such as publication or outcome reporting bias,\(^3\) sponsors\(^4\),\(^5\) and author’s financial conflict of interests,\(^6\),\(^7\) or data falsification or fabrication\(^8\),\(^9\) just to name a few. Secondly, because they read all that has been published worldwide on a given topic, systematic reviewers are in a very good position to identify malpractices or even suspect new cases of misconduct.\(^10\) For instance, systematic reviewers can detect “copy-paste like” duplicate publications,\(^11\),\(^12\) and it has been suggested that they should also try to identify trials lacking reports of formal ethical approval.\(^13\),\(^14\) In some cases, systematic reviewers may come to suspect data fabrication as well, although there is to date no widely accepted and validated way to detect such cases.

### Box 3

**The Fujii case: uncovering fraud by a systematic review**

Prof Yoshitaka Fujii was a Japanese anaesthesiologist. His area of research focused on the prevention of post-operative nausea and vomiting. He worked in various institutions in Japan. Suspicion regarding the reliability of his research arose in 2001 with the publication of a systematic review investigating the impact of a drug “granisetron” on the incidence of post-operative nausea and vomiting by Peter Kranke et al.\(^15\) In this systematic review, it became obvious that something was wrong with the Fujii papers; authors of the systematic review realised that the incidences of headaches in the control groups of these trials were too similar to have happened by random chance. However, they were unable to prove data falsification and shared their concern through a letter to the editor.\(^16\) In 2012, Toho University asked for the retraction of 8 of Prof Fujii’s trials for “failure to seek formal ethical approval”. This was followed by the publication of a report by John Carlisle\(^17\) who analysed the data reported in Fujii’s articles and showed that the data were not consistent with what would be expected on real samples of humans (lack of variability of different variables). An investigation was then launched, and ended with the dramatic conclusion that, of 212 articles examined, 126 had been totally fabricated, 46 contained fabricated data, for 37 no conclusion could be drawn. Only three trials were valid.\(^18\)

The problem is that it remains unclear to many systematic reviewers what the correct procedure should be when uncovering these misconducts;\(^19\) this will be discussed in chapter 5.5. Finally systematic reviewers can serve to guide and improve the design and relevance of future research. For example, they may identify when a research question has become irrelevant, because the answer to that question is known, and highlight exactly what still needs to be investigated in order to avoid future useless research.\(^20\)
TABLE 2
Systematic reviews in the context of research malpractice

<table>
<thead>
<tr>
<th>Potential malpractices</th>
<th>Roles of the systematic review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malpractices that do not threaten the scientific knowledge or the conclusions of systematic reviews</td>
<td></td>
</tr>
<tr>
<td>Incomplete review of existing literature</td>
<td>Can serve as a basis for identification of future research needed</td>
</tr>
<tr>
<td>Guest authorship</td>
<td>Can identify plagiarism</td>
</tr>
<tr>
<td>Plagiarism</td>
<td>Can identify trials that fail to report on formal ethical approval or patient consent</td>
</tr>
<tr>
<td>Failure to submit the protocol to ethics committee for approval</td>
<td>Can identify trials whose protocol is not accessible</td>
</tr>
<tr>
<td>Failure to seek patient informed consent</td>
<td></td>
</tr>
<tr>
<td>Failure to register protocol</td>
<td></td>
</tr>
<tr>
<td>Malpractices that can threaten the scientific knowledge, but not necessarily the conclusions of systematic reviews</td>
<td></td>
</tr>
<tr>
<td>Clinically irrelevant or unclear research question</td>
<td>Can identify questions that have already been answered, and those who deserve further investigation.</td>
</tr>
<tr>
<td>Wrong design</td>
<td>Can exclude inadequate designs from analyses</td>
</tr>
<tr>
<td>Deviation from protocol</td>
<td>Can identify protocol deviation and change in primary outcome if protocol accessible</td>
</tr>
<tr>
<td>Changing the primary outcome</td>
<td>Can recalculate measures of effect</td>
</tr>
<tr>
<td>Lack of analysis plan or wrong statistical tests applied</td>
<td>Can pool underpowered trials</td>
</tr>
<tr>
<td>Lack of sample size calculation</td>
<td>Can rate the risk of bias through critical appraisal</td>
</tr>
<tr>
<td>Selection or information bias</td>
<td>Can contact study authors to clarify outcome definition and unearth unreported outcome</td>
</tr>
<tr>
<td>Violation of blinding</td>
<td>Not considered by systematic reviewers</td>
</tr>
<tr>
<td>Ill-defined primary and secondary outcome</td>
<td>Not considered for quantitative systematic reviews</td>
</tr>
<tr>
<td>Selective reporting of outcomes</td>
<td>Can search for unpublished trials or test for publication bias</td>
</tr>
<tr>
<td>Selective or self-citation</td>
<td>Can identify duplicate publications and exclude duplicates from analyses</td>
</tr>
<tr>
<td>Biased or over-interpretation of results</td>
<td></td>
</tr>
<tr>
<td>Failure to publish research</td>
<td></td>
</tr>
<tr>
<td>Duplicate publication</td>
<td></td>
</tr>
<tr>
<td>Malpractice that threatens the scientific knowledge and the conclusions of systematic reviews</td>
<td></td>
</tr>
<tr>
<td>Data falsification (Ignoring outliers)*</td>
<td></td>
</tr>
<tr>
<td>Data fabrication*</td>
<td></td>
</tr>
<tr>
<td>Not reporting protocol violation</td>
<td></td>
</tr>
<tr>
<td>Not declaring conflicts of interest</td>
<td></td>
</tr>
<tr>
<td>Ghost authorship (with conflict of interest)</td>
<td></td>
</tr>
<tr>
<td>Peer-review organisation</td>
<td></td>
</tr>
<tr>
<td>Non-correction of identified errors</td>
<td></td>
</tr>
<tr>
<td>Non-retraction of articles that warrant retraction</td>
<td></td>
</tr>
</tbody>
</table>

*May be suspected in exceptional cases although inconsistently.
5 Personal contributions in this area of research

The following chapter reports my personal contributions regarding the problems of research malpractices and misconduct for systematic reviewers.

Chapter 5.1 is an example of a systematic review performed to evaluate the impact of adding ketamine to an anaesthetic-analgesic regimen in order to achieve a better control of postoperative pain. This article illustrates how the procedures underlying systematic reviewing routinely lead to the identification of common errors, suboptimal reporting of research outcomes, but also how systematic reviewers work to “fill in the gaps”, to correct poor quality or badly reported research. Chapter 5.2 explores the question of an overlooked threat to systematic reviews: the un-retraction of articles that warranted retraction. Although some articles had been recognised as fraudulent and required to be retracted, the retraction process was not straightforward and some articles may remained un-retracted, and their status unknown to systematic reviewers. Chapter 5.3 is a re-analysis of systematic reviews that illustrates what may happen to the conclusions of systematic reviews when fraudulent data has been inadvertently included into meta-analyses. Chapter 5.4, is a cross-sectional analysis of systematic reviews and survey of their authors that examines the hypothesis that systematic reviewers are in a privileged position to identify or suspect research malpractices or misconduct, and that they are both aware of this, and ready to apply protective measures accordingly. Finally chapter 5.5 proposes a new role for systematic reviewers, which consists of guiding researcher to avoid performing research on clinically irrelevant questions, and helping them improve their study designs.
5.1 Ketamine for post-operative pain - a quantitative systematic review of randomised trials

Authors: Elia N and Tramèr MR
Published in: Pain 2005; 113: 61-70
IF 2005: 4.309

Summary

In this systematic review with meta-analyses, our aim was to identify the impact of adding ketamine to anaesthesia for the control of postoperative pain and adverse effects. We included 53 trials (about 3000 patients) originating from 25 countries, and reporting on many different settings.

This example illustrates that systematic reviewing implies searching for published and unpublished trials (through contacts with the manufacturer of ketamine), in any language (retrieved articles were written in English, German, Spanish and Japanese), contacting study authors to search for additional information (11 of 20 authors contacted answered) and the rating of the trial quality (which was performed, ten year ago, using a modified Oxford scale based on randomisation, concealment of allocation, blinding and reporting of the flow of patients).

This example also illustrates how systematic reviewers may identify studies with inadequate control groups, (69 trials identified, and exclude from analyses), identify problems related to selective non-reporting of adverse effects, of reporting of non-relevant outcomes, deplore the small sizes of the trials performed in anaesthesiology and the fact that some authors are unable or not willing to respond to inquiries.

Finally, this systematic review illustrates how the lack of a rational clinical research agenda can result in unclear evidence of the effect of an intervention despite randomisation of about 3000 patients, and provides an important, useful and clear research agenda to guide further research.

The interested reader may want to read a recent systematic review from the same authors testing the impact of Ketamine in a very specific setting (mixed with an opioid in a patient-controlled device) on postoperative pain and side effects. Although the topic and the authors remained the similar, the methods used have changed, especially regarding the evaluation of the quality of trials, and the use of trial sequential analyses. This article was not included in the present dissertation due to space limitation.

Ref: Benefit and harm of adding ketamine to an opioid in a patient-controlled analgesia device for the control of postoperative pain: systematic review and meta-analyses of randomized controlled trials with trial sequential analyses. Assouline B, Tramèr MR, Kreienbühl L, Elia N PAIN (in press)
Ketamine and postoperative pain – a quantitative systematic review of randomised trials

Nadia Elia*, Martin R. Trame`r

EBCAP Institute (Evidence-Based Critical care, Anaesthesia and Pain treatment), Division of Anaesthesiology, Geneva University Hospitals, 24 Rue Micheli-du-Crest, CH-1211 Geneva 14, Switzerland

Received 5 May 2004; received in revised form 10 September 2004; accepted 28 September 2004

Abstract

Ketamine, an N-methyl-D-aspartate receptor antagonist, is known to be analgesic and to induce psychomimetic effects. Benefits and risks of ketamine for the control of postoperative pain are not well understood. We systematically searched for randomised comparisons of ketamine with inactive controls in surgical patients, reporting on pain outcomes, opioid sparing, and adverse effects. Data were combined using a fixed effect model. Fifty-three trials (2839 patients) from 25 countries reported on a large variety of different ketamine regimens and surgical settings. Sixteen studies tested prophylactic intravenous ketamine (median dose 0.4 mg/kg, range (0.1–1.6)) in 850 adults. Weighted mean difference (WMD) for postoperative pain intensity (0–10 cm visual analogue scale) was −0.89 cm at 6 h, −0.42 at 12 h, −0.35 at 24 h and −0.27 at 48 h. Cumulative morphine consumption at 24 h was significantly decreased with ketamine (WMD −15.7 mg). There was no difference in morphine-related adverse effects. The other 37 trials tested in adults or children, prophylactic or therapeutic ketamine orally, intramuscularly, subcutaneously, intra-articulary, caudally, epidurally, transdermally, peripherally or added to a PCA device; meta-analyses were deemed inappropriate. The highest risk of hallucinations was in awake or sedated patients receiving ketamine without benzodiazepine; compared with controls, the odds ratio (OR) was 2.32 (95% CI, 1.09–4.92), number-needed-to-harm (NNH) 21. In patients undergoing general anaesthesia, the incidence of hallucinations was low and independent of benzodiazepine premedication; OR 1.49 (95% CI 0.18–12.6), NNH 286. Despite many published randomised trials, the role of ketamine, as a component of perioperative analgesia, remains unclear.

© 2004 International Association for the Study of Pain. Published by Elsevier B.V. All rights reserved.

Keywords: Systematic review; Meta-analysis; Surgery; Postoperative; Pain; Analgesia

1. Introduction

The identification of the N-methyl-D-aspartate (NMDA) receptor and its role in pain perception has brought a new interest to ketamine as a possible adjuvant to multimodal pain treatment (Klepstad et al., 1990). During the last 15 years, a large number of clinical trials have been published that tested this NMDA-receptor antagonist for the management of postoperative pain. However, the evidence for a beneficial effect of ketamine in this setting remains controversial, and the fear of psychomimetic adverse effects such as hallucinations or nightmares limits its widespread use in clinical practice.

In 1999, Schmid et al. published a review on ketamine for postoperative pain treatment (Schmid et al., 1999). The authors concluded that ketamine was analgesic and useful in surgical patients. However, the reviewed trials reported on a wide diversity of clinical settings and ketamine regimens, and the authors were unable to quantify the analgesic effect and the risk of harm of ketamine. Many new reports have been published since. Also, to aid in clinical decision making, health care providers need to know how well ketamine works as an analgesic and what the incidence of adverse effects is. Thus, valid quantitative information is needed about minimal effective dose, dose-responsiveness,
and adverse effect profile. The aim of this systematic review is to address these issues.

2. Methods

We followed the QUOROM statement that recommends standards to improve the quality of reporting of meta-analyses (Moher et al., 1999).

2.1. Systematic search

Published reports of randomised trials testing ketamine for postoperative pain management were searched in Medline, Embase, CINHAL, Biosis previews, Indmed, and the Cochrane Controlled Trials Register. We used the free text and MeSH terms ‘ketamine’, ‘ketalar’, ‘ketanest’, ‘keta*’, ‘surgery’, ‘surgical’, ‘post-surgical’, ‘pain treatment’, ‘analgiesia’ and ‘analgesic’, without language restriction. We limited our search to randomised trials and humans. The last electronic search was in November 2003. We identified additional studies from the bibliographies of retrieved reports and reviews (Eide et al., 1995; Granny et al., 2000; Kohrs and Durieux, 1998; Schmid et al., 1999; White et al., 1982). We contacted the manufacturer of ketamine (Pfizer AG, 8052 Zürich, Switzerland) and asked for additional reports including unpublished data. We contacted authors of original reports for translation or to obtain additional information.

2.2. Inclusion and exclusion criteria

We considered randomised trials with an inactive (placebo or ‘no treatment’) control group, and that reported on postoperative pain outcomes or adverse drug reactions related to ketamine. Relevant pain outcomes included pain intensity, time to first analgesic request, postoperative cumulative morphine consumption and morphine-related adverse effects. We excluded trials including less than 10 patients per group and those reporting on premedication, chronic pain, or emergency medicine. Data from animal studies, abstracts, letters or reviews were not considered.

2.3. Validity scoring

One author (NE) screened the abstracts of all retrieved reports and excluded articles that clearly did not meet our inclusion criteria. Both authors then independently read the included reports and assessed their methodological validity using a modified 7-point 4-item Oxford scale (Box 1) (Jadad et al., 1996; Pasquina et al., 2003). We resolved discrepancies by discussion. As there was a prior agreement that we would exclude reports without randomisation, the minimum score of an included trial was one, and the maximum score was seven.

2.4. Data extraction

We extracted information about ketamine (time and route of administration, dose), number of patients enrolled and analysed, observation period, surgery, postoperative analgesia and pain outcomes. Dichotomous data on presence or absence of adverse effects were extracted. There was a pre-hoc agreement that a morphine-sparing effect would be clinically relevant if there was a decrease in morphine-related adverse effects. Definitions of adverse effects were taken as reported in the original trials.

To facilitate comparisons between trials, we normalised ketamine regimens to milligrams per kilogram of bodyweight (mg/kg). When the trials did not report on average bodyweight of the studied population, we assumed it was 70 kg.

Ketamine administration at induction of anaesthesia and during or immediately after surgery (before the patient complained of pain) was regarded as a prophylactic regimen. Postoperative administration in a patient with overt pain was defined as a therapeutic regimen. When a trial tested ketamine before versus after surgery, and there was no evidence of any difference, we combined the data.

2.5. Meta-analyses

For continuous data, we calculated weighted mean differences (WMD) with 95% confidence interval (CI). If the authors did not report on mean values and standard deviations, we contacted them. If they did not answer and the data were reported as graphs, we extracted the data from the graphs.

Dichotomous data on adverse effects were summarised using relative risks (RR) with 95%CI. For rare outcomes (for instance, psychomimetic adverse effects) we computed Peto odds ratios (OR) that deal better with zero cells. If the 95%CI included one, we assumed that the difference between ketamine and control was not statistically significant. To estimate the clinical relevance of an adverse effect we calculated the number-needed-to-harm (NNH)
(McQuay and Moore, 1998). NNH were calculated whether or not the OR suggested a statistically significant difference between ketamine and control. We used a fixed effect model throughout since we combined data only when they were clinically homogeneous. Analyses were performed using ReviewManager software (version 4.0, Cochrane Collaboration) and Excel. Data were graphically plotted using forest plots to evaluate treatment effects, and L’Abbé plots (event rate scatters) to explore variability of the data.

3. Results

3.1. Systematic search

We identified 241 trials; 188 were subsequently excluded (Appendix A: Flow chart). Fifty-three valid randomised trials tested ketamine in 2,721 adults and children (references, see included randomised trials; detailed information, see Appendix A: Analysed trials). The manufacturer of ketamine did not provide any data. We contacted the authors of 20 reports to obtain additional information; eleven answered (Ngan Kee et al., 1997; Lee and Sanders, 2000; Kakinohana et al., 2000; Burstal et al., 2001; Subramaniam et al., 2001a; Unlügenc¸ et al., 2002; Jakšch et al., 2002; Unlügenc¸ et al., 2003; Rosseland et al., 2003; Kararmaz et al., 2003; Xie et al., 2003). We did not retrieve any unpublished data. The reports were published between 1971 and 2003. They were written in English (n = 49), German (2), Spanish (1), and Japanese (1), and originated from 25 countries (six from Turkey; five from the US; four each from China and Germany; three each from France, Norway and the UK; two each from Australia, Belgium, Brazil, Greece, India, Israel and Japan; and one each from Austria, Denmark, Ireland, New Zealand, Saudi Arabia, South Africa, South Korea, Spain, Sweden, Taiwan and United Arab Emirates). Two studies declared sponsorship by the manufacturer (Reeves et al., 2001; Weber and Wulf, 2003).

The median number of patients receiving ketamine per study was 25 (range, 10–105). The median modified Oxford scale was 4 (range, 2–6). Four trials tested S(+)-ketamine (Himmelseher et al., 2001; Lauretti et al., 2001; Jakšch et al., 2002; Weber and Wulf, 2003), all others used racemic ketamine. The trials described a large variety of ketamine regimens and surgical settings (Table 1). Two trials tested two different regimens (De Kock et al., 2001; Xie et al., 2003). The largest clinically homogenous subgroup (16 trials) tested prophylactic intravenous ketamine in adults undergoing general anaesthesia. These trials shall be discussed in more detail.

3.2. Prophylactic intravenous ketamine

3.2.1. Trials

Of the 16 trials, two only (Ngan Kee et al., 1997; Roytblat et al., 1993) had been included in the review by Schmid et al. (1999). However, of those considered by Schmid et al., seven were excluded by us: four lacked an inactive control group (Fu et al., 1997; Jahangir et al., 1993; Owen et al., 1987; Wilder-Smith et al., 1998), one was on healthy volunteers (Maurset et al., 1989), one had less than 10 patients per group (Tverskoy et al., 1994), and one was not randomised (Clausen et al., 1975).

In the 16 trials, 889 adults were randomised and data from 850 were subsequently analysed by the original investigators. Most trials reported on follow-up periods of 24–48 h. One trial reported on additional outcomes after three days (Menigaux et al., 2001), one after five days (Jakšch et al., 2002), and one after one year (De Kock et al., 2001).

3.2.2. Surgical settings and ketamine regimens

Surgeries were abdominal (7 trials), gynaecologic (4), orthopaedic (3), maxillo-facial (1), and outpatient surgery (1). In 11 trials, a single bolus of ketamine was injected at induction of anaesthesia or immediately after surgery. In four, a preoperative bolus was followed by a continuous infusion throughout surgery (De Kock et al., 2001; Guignard et al., 2002; Jakšch et al., 2002; Kararmaz et al., 2003). One trial tested both bolus injection and continuous infusion (Heinke and Grimm, 1999). In three trials, an identical dose of ketamine was administered either before surgery or at the end of surgery; in none of these did the time of administration have an impact on the outcome (Dahl et al., 2000; Gilabert Morell and Sanchez Perez, 2002; Menigaux et al., 2000). The average body weight in 51 trials was 70.3 kg; the median dose of ketamine across all trials was 0.4 mg/kg (range, 0.1–1.6).

### Table 1

<table>
<thead>
<tr>
<th>Clinical setting</th>
<th>Nb trials</th>
<th>Nb patients</th>
<th>Regimens</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Route</strong></td>
<td><strong>Goal</strong></td>
<td><strong>K/C</strong></td>
<td><strong>Mg/kg</strong></td>
</tr>
<tr>
<td>Adults</td>
<td>IV</td>
<td>P</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>IV PCA</td>
<td>T</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Epid</td>
<td>P</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Epid</td>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>IV &gt; 24 h</td>
<td>P + T</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>IV (sedation)</td>
<td>P</td>
<td>3</td>
</tr>
<tr>
<td>Children</td>
<td>IV</td>
<td>P</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Caudal</td>
<td>P</td>
<td>4</td>
</tr>
<tr>
<td>Other regimens in adults or children</td>
<td>PO, IM, IA, IV, P/T</td>
<td>10</td>
<td>249/195</td>
</tr>
</tbody>
</table>

**Clinical setting** includes abdominal, gynaecologic, orthopaedic, maxillo-facial, and outpatient surgery. **Route** includes IV, epidural, PO, per os; IM, intramuscular; IA, intra-articular; SC, subcutaneous; TTS, transcutaneous transport system; IV PCA, patient controlled analgesia; Epid, epidural; P, prophylactic; T, therapeutic; K, ketamine; C, control. Regimens have been converted to (cumulative) mg/kg taking into account average bodyweights as reported in the original trials. Ketamine regimens are expressed as median [range].

**Number of trials do not add up since 2 trials tested two routes of administration, IV, intravenous; Epid, epidural; PO, per os; IM, intramuscular; IA, intra-articular; SC, subcutaneous; TTS, transcutaneous transport system; NB, nerve block (brachial plexus and Bier’s block); PCA, patient controlled analgesia; P, prophylactic; T, therapeutic; K, ketamine; C, control. Regimens have been converted to (cumulative) mg/kg taking into account average bodyweights as reported in the original trials. Ketamine regimens are expressed as median [range]."
3.2.3. Analgesic efficacy of intravenous ketamine

3.2.3.1. Pain intensity. Ten trials reported on pain intensity at rest, measured on a 0–10 cm visual analogue scale (Fig. 1; detailed data see Appendix A: VAS of Pain Intensity). In controls, pain intensity rarely reached more than 4 cm on the 10 cm scale. Combined data showed a consistent and statistically significant decrease in pain intensity at rest with ketamine compared with control at 6 h postoperatively (WMD, −0.89 cm), 12 h (−0.42 cm), 24 h (−0.35 cm), and 48 h (−0.27 cm). We found no evidence of a relationship between the dose of ketamine and analgesic efficacy. Four trials reported on pain intensity on movement (De Kock et al., 2001; Jaksch et al., 2002; Kararmaz et al., 2003; Ngan Kee et al., 1997), each at different time points; meta-analysis was deemed inappropriate.

3.2.3.2. Cumulative morphine consumption at 24 h. Seven trials reported on morphine consumed with a PCA pump during the first 24 h after surgery. In two trials (De Kock et al., 2001; Guignard et al., 2002) the data was not reported as means ± standard deviations and the authors did not respond to our inquiries. In four trials, average morphine consumption in controls was between 34.2 and 49.7 mg. In one trial, ketorolac was added to the morphine in the PCA; in that trial, average cumulative morphine consumption in controls was 15.6 mg only (Gilabert Morell and Sanchez Perez, 2002). All trials except one (Jaksch et al., 2002) reported on a statistically significant decrease in morphine consumption with ketamine; morphine-sparing ranged from 9% (Jaksch et al., 2002) to 47% (Menigaux et al., 2000) (median: 32%). When data from all four trials were combined, WMD in favour of ketamine was about −16 mg. Including the trial that was using ketorolac did not change that result. (Fig. 2A; detailed data see Appendix A: Cumulative Morphine Consumption).

3.2.3.3. Opioid-related adverse effects. In 12 trials, postoperative analgesia was with systemic morphine, piritramide or fentanyl; two of those did not report on opioid-related adverse effects (Heinke and Grimm, 1999; Ngan Kee et al., 1997). One study reported on the number of patients...
presenting less than five episodes of nausea (De Kock et al., 2001); these data could not be combined with nausea data from the other trials. The nine remaining trials reported on presence or absence of nausea (5 trials), vomiting (4), nausea or vomiting (4), pruritus (2), drowsiness (2), or urinary retention (3) (Fig. 2B; detailed data see Appendix A: Opioid-Related Adverse Effects). Combined data showed no statistically significant difference between ketamine and control for any of these opioid-related adverse effects.

3.2.3.4. Time to first request of analgesic. Seven trials reported on the time when patients first requested an analgesic postoperatively (see Appendix A: Time to First Request of Analgesic). In six of seven trials, average delay in controls ranged from 10 to 30 min. In one study, wound infiltration with ropivacaine 200 mg increased this delay to 121 min (Papaziogas et al., 2001). When data from all seven trials were combined, the average improvement with ketamine was about 16 min. Exclusion of the trial that was using ropivacaine did not change that result.

3.2.3.5. Long term follow up. In one trial, patients were contacted after two weeks, one month, six months and one year and were asked if they felt any pain or particular sensations at the scar area, and whether this pain required medication (De Kock et al., 2001). At six months, significantly fewer patients who had received ‘high dose’ intravenous ketamine (bolus 0.5 mg/kg, infusion 0.25 mg/kg/h during surgery) suffered from residual pain requiring analgesic medication. After one year, three of 17 controls still had pain compared with none of those who had received intravenous ketamine.

3.3. Other ketamine regimens

Thirty-seven trials tested a variety of regimens and routes of administration of ketamine in children or adults (Appendix A: Analysed Trials).

3.3.1. Ketamine added to an opioid in an intravenous PCA device

In five trials (284 adults), ketamine was added to morphine or tramadol in a PCA. Regimens ranged from 1 to 2.5 mg of ketamine per mg of morphine, and 0.2 mg of ketamine per mg of tramadol. Meta-analysis was deemed inappropriate. Authors of three trials concluded that ketamine decreased postoperative pain intensity; the two others concluded that it did not.

3.3.2. Epidural ketamine in adults

Ten trials reported on prophylactic epidural ketamine in 454 adults. Two further trials reported on therapeutic
epidural ketamine in 131 patients. There was a large variability in epidural regimens; meta-analysis was deemed inappropriate. Authors of five of ten prophylactic and of both therapeutic studies concluded that ketamine decreased postoperative pain; the five others concluded that it did not.

3.3.3. Prolonged (> 24 h) intravenous ketamine in adults

In three trials (112 patients), an intravenous infusion of ketamine was started at the beginning of surgery and continued until 24 or 48 h after surgery. Meta-analysis was deemed inappropriate. Authors of one report concluded that ketamine decreased postoperative pain; the two others concluded that it did not.

3.3.4. Intravenous ketamine in sedated adults

Three trials reported on intravenous ketamine during sedation in 36 intubated patients in the intensive care unit, and 166 patients undergoing breast biopsy or cataract extraction. Surgical procedures were not comparable and pain measurements were inconsistent; meta-analysis was deemed inappropriate. Authors of all reports concluded that ketamine was beneficial during sedation since it decreased the number of rescue analgesics that were needed during surgery and the number of patients with inadequate sedation.

3.3.5. Intravenous and caudal ketamine in children

Two paediatric trials (76 children) tested prophylactic intravenous ketamine, and four (168 children) reported on ketamine added to caudal ropivacaine, bupivacaine, or alfentanil. Postoperative pain measurements were inconsistent; meta-analysis was deemed inappropriate. All authors concluded that ketamine decreased postoperative pain.

3.3.6. Other ketamine regimens

Ten trials (444 patients) reported on a variety of other regimens. Ketamine was given intra-articularly, intramuscularly, subcutaneously, orally, transdermally, added to a brachial plexus block with local anaesthetic, as an adjuvant to intravenous regional anaesthesia, or as a combination of an intravenous preoperative bolus with postoperative PCA. Authors of five reports concluded that ketamine was beneficial; the five others concluded that it was not.

3.4. Safety analysis

3.4.1. Hallucinations

Thirty trials reported on hallucinations. We tested the impact of the dose of ketamine, of a premedication with benzodiazepines, and of patients’ vigilance at the time of drug administration on the risk of hallucinations. Benzodiazepines were midazolam, diazepam, lorazepam, lormetazepam, or temazepam. We assumed that for these sensitivity analyses, the route of administration of ketamine was not a factor.

There was no graphical evidence of a relationship between the dose of ketamine and the risk of having hallucinations (Appendix A: Risk of Hallucinations According to the Weight Adjusted Ketamine Dose).

To test the impact of concomitant benzodiazepines and patients’ vigilance we performed four subgroup analyses (Fig. 3). In eight trials, patients received ketamine during general anaesthesia after benzodiazepine premedication (Abdel-Ghaffar et al., 1998; De Kock et al., 2001; Roytblat et al., 1993; Gilabert Morell and Sanchez Perez, 2002; Guignard et al., 2002; Papaziogas et al., 2001; Subramaniam et al., 2001b; Weir et al., 1998). Three of 289 (1%) patients experienced hallucinations with ketamine, compared with one of 154 (0.6%) controls; OR 1.49 (95% CI, 0.18–12.6), NNH 257. In four trials, 107 patients received ketamine (91 controls) during general anaesthesia without a benzodiazepine premedication (Lee and Sanders, 2000; Menigaux et al., 2000; Menigaux et al., 2001; Ozbek et al., 2002). There were no reports of hallucinations. Thus, in patients undergoing a general anaesthetic with or without benzodiazepine premedication, three of 396 (0.8%) presented hallucinations with ketamine compared with one of 245 (0.4%) controls; OR 1.49 (95% CI, 0.19–12.6), NNH 286.

Fig. 3. Risk of hallucinations according to patients’ vigilance and premedication with benzodiazepines. Each symbol represents one comparison between ketamine and control. GA, general anaesthesia; benzo, benzodiazepine; NNH, number-needed-to-harm. Benzodiazepines were midazolam, diazepam, lorazepam, lormetazepam or temazepam.
In seven trials (eight comparisons), patients were awake or sedated after a benzodiazepine premedication (Adriaenssens et al., 1999; Azevedo et al., 2000; Badrinath et al., 2000; Ilkjaer et al., 1998; Lauretti et al., 2001; Subramaniam et al., 2001a; Unlüügen et al., 2002). Of 217 patients who received ketamine, nine (4.1%) had hallucinations, compared with two of 155 (1.3%) controls; OR 2.19 (95%CI, 0.58–8.27), NNH 35. In nine trials, patients were awake or sedated without a benzodiazepine premedication (Frey et al., 1999; Gorgias et al., 2001; Hercock et al., 1999; Huang et al., 2000; Joachimsson et al., 1986; Lee et al., 2002; Qureshi et al., 1995; Reeves et al., 2001; Yanli and Eren, 1996). Of 230 patients who received ketamine, 24 (10.4%) had hallucinations, compared with 11 of 193 (5.7%) controls; OR 2.32 (95%CI, 1.09–4.92), NNH 21. Thus, of all awake or sedated patients, 33 of 447 (7.4%) presented hallucinations with ketamine compared with 13 of 348 (3.7%) controls; OR 2.28 (95%CI, 1.19–4.40), NNH 27.

3.4.2. Nightmares

Thirteen trials reported on the incidence of nightmares. With ketamine, 10 of 421 (2.4%) patients had nightmares, compared with two of 262 (0.8%) controls; OR 2.64 (95%CI, 0.76–9.12), NNH 62.

3.4.3. Pleasant dreams

Eight trials reported on pleasant dreams. With ketamine, 29 of 159 patients (18.2%) had pleasant dreams, compared with 14 of 145 (9.7%) controls; OR 1.96 (95%CI, 1.01–3.81), NNT 12.

3.4.4. Visual disturbances

Thirteen trials reported on visual disturbances (diplopia or nystagmus). With ketamine, 23 of 373 (6.2%) patients had visual disturbances, compared with seven of 271 (2.6%) controls; OR 2.34 (95%CI, 1.09–5.04), NNH 28.

4. Discussion

4.1. Efficacy and harm

Knowing that ketamine has some analgesic effect in surgical patients begs the question as to whether it should be used more frequently as an adjuvant to multimodal perioperative analgesia. When administered intravenously during anaesthesia in adults, ketamine decreases postoperative pain intensity up to 48 h, decreases cumulative 24 h morphine consumption, and delays the time to first request of rescue analgesic. When assessing the clinical relevance of these potentially beneficial effects, several issues need to be considered. Firstly, pain intensity was decreased by about 1 cm on a 10 cm pain scale at 6 h; at subsequent time points, this benefit further decreased but the effect was still statistically significant at 48 h. In control groups, pain intensity was about 4 cm on the 10 cm scale; thus, patients who received ketamine had a decline in pain intensity of about 25% at 6 h and of about 20% at 24 h compared with those who received a placebo. Since patients in control groups did not experience very severe pain, the clinical relevance of this improvement remains unclear (Kalso et al., 2002). Secondly, in clinical practice, both decreased pain intensity and decreased opioid consumption may be regarded as two linked proxies of analgesic efficacy, and it may not be feasible to clearly separate between these two endpoints. During the first 24 postoperative hours, controls consumed about 40 mg of morphine on average; this corresponds to the usual average amount of morphine consumed during the first day after major surgery (Walder et al., 2001). In four of five trials, this dose of morphine was reduced by 27–47%, and at the same time, patients had a reduction in pain intensity. Thus, the beneficial albeit moderate effect of ketamine on pain intensity should be interpreted in conjunction with the opioid-sparing effect. Thirdly and interestingly, there was no decrease in the incidence of morphine-related adverse effects although morphine consumption was clearly reduced with ketamine. One reason may be that these trials mainly concentrated on efficacy and did not systematically evaluate and report on adverse effects. This limitation on the reporting of adverse effects has been described in other pain settings (Edwards et al., 1999). Another reason may be that the decrease in morphine consumption was not strong enough to impact the incidence of morphine-related adverse effects; this would weaken the clinical relevance of the opioid-sparing. Finally, with ketamine, the delay until patients requested the first rescue analgesic was prolonged by about 16 min. This result was statistically significant but of no clinical importance; however, it was consistent with the overall analgesic efficacy of ketamine. One trial only looked at long term outcomes (De Kock et al., 2001). Although that study included a limited number of patients, the authors were able to show a beneficial effect of ketamine on the persistence of painful sensations around the scar for up to 6 months after surgery. These data support the biological basis of ketamine as an antagonist at the NMDA receptor.

In children, there was some evidence for an analgesic effect with intravenous or caudal ketamine. All paediatric trials concluded that ketamine may be of use. The role of ketamine as an adjunct to morphine-PCA, however, remains unclear.

A further issue that needs to be considered when discussing the usefulness of ketamine as a component of perioperative analgesia is harm. The risk of psychomimetic adverse effects such as hallucinations is perhaps the main reason why many clinicians are apprehensive in using ketamine. Sensitivity analyses provided some evidence that patients’ vigilance, and less so concomitant use of benzodiazepines, had an impact on the risk of hallucinations. Three conclusions may be drawn. Firstly, in anaesthetised patients, the risk of ketamine-induced hallucinations is minimal. Secondly, benzodiazepines should not be regarded...
as an universally effective protection against hallucinations; even with benzodiazepine premedication, one in 35 non-anaesthetised patients will have hallucinations who would not have done so had they not received ketamine. Finally, ketamine should be given very carefully to patients who are not anaesthetised; patients should be informed about this potential risk and they should be carefully observed during the procedure. There were also other ketamine-related adverse effects described such as pleasant dreams or visual disturbances; the clinical relevance of those is not obvious.

4.2. Limitations

This systematic review has several limitations; most of them are related to methodological weaknesses of the original studies. For example, the original trials did not allow to identify patient-related predictive factors for the emergence of hallucinations or to distinguish between the efficacy of different benzodiazepines. Also, we were unable to establish a dose-response for efficacy or harm, nor were we able to quantify the impact of perioperative ketamine on the emergence of chronic pain. One of the major problems was the wide variability in clinical settings, ketamine regimens and reported outcomes. Fifty-three trials were conducted in 25 countries. This illustrates a large and worldwide interest in this drug. However, most trials were of limited size: groups rarely exceeded 30 patients. Also, for some routes of administration, such as transdermally or intra-articularly, distinct pharmacokinetic profiles made it impossible to combine these trials with the others. There was a large variety in outcome measures, and those outcomes that might be regarded as clinically relevant (i.e. opioid-related adverse effects) were inconsistently reported. Finally, data reporting was often unsatisfactory; multiple trials presented data as figures only and authors, when asked for more detailed information, were unable or unwilling to respond to our enquiry. As a consequence, for the majority of the regimens, data pooling was impossible or was deemed inadequate. All these problems have been described before for other acute pain settings (Walder et al., 2001). It is tempting to believe that this poor output, despite a large number of randomised patients, was mainly due to the obvious lack of a rational clinical trial program. Indeed, two trials only declared sponsorship by the manufacturer, suggesting that ketamine has never reached high priority for the manufacturer. Ketamine for chronic pain (Hocking and Cousins, 2003) and for cancer pain (Bell et al., 2003) has been reviewed recently. The authors concluded that the evidence for the efficacy of ketamine in the chronic pain setting was weak due to a lack of high quality, randomised trials with an appropriate number of patients and standardised ketamine regimens.

4.3. Research agenda

This systematic review highlights some areas where further research with ketamine is warranted but also areas where further research is not justified. For instance, three trials were unable to show a benefit when ketamine was administrated preemptively, echoing similar negative results of preemptive analgesia with other agents (Moiniche et al., 2002). We were unable to identify a dose-response for analgesic efficacy, and in anaesthetised patients, the risk of psychomimetic adverse effects was not a limiting factor. Thus, future studies could test, in anaesthetised patients, higher doses than those reported in these trials. Patients should undergo major or painful procedures; only then, the analgesic efficacy of ketamine may be measured (Kalso et al., 2002). Trials should include long-term follow-up (up to one year) to investigate the impact of perioperative ketamine on chronic pain. Further studies should be of reasonable size, and should report on clinically relevant endpoints of both analgesic efficacy (VAS of pain intensity, 24 h cumulative morphine consumption) and harm (psychomimetic effects, prolonged sedation). A further interesting setting that should be studied is that of morphine-resistant pain. In postoperative patients who experienced only partial pain relief with morphine (i.e. VAS pain intensity > 6/10 despite intravenous morphine 0.1 mg/kg), small doses of ketamine (0.25 mg/kg, up to three times) rapidly improved pain relief (Weinbroum, 2003). These observations support the role of the NMDA receptor in the acute pain setting. Adverse drug reactions should be carefully watched for and any alleged opioid-sparing should be confirmed by a reduction in opioid-related adverse effects. Finally, before administering ketamine to patients who are not anaesthetised, a careful evaluation of the expected benefits and risks should be established.

Conflict of interest statement

None declared

Acknowledgements

Special thanks go to D. Haake for his help in searching electronic databases, to G. Elia for statistical help, and to Drs R. Burstal, W. Jakusch, M. Kakihana, A. Kararmaz, H.M. Lee, W.D. Ngan Kee, L.A. Rosseland, K. Subramaniam, H. Unlüenç, and H. Xie, who responded to our inquiries. Dr Tramèr was a beneficiary of a PROSPER grant (No 3233-051939.97/2) from the Swiss National Science Foundation. The salary of Dr Elia was provided by the Swiss National Science Foundation (No 3200-064800.01/1) and by the EBCAP Institute.

Appendix A

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.pain.2004.09.036
References


Included randomised trials


N. Elia, M.R. Tramèr / Pain 113 (2005) 61–70
5.2 Fate of articles that warranted retraction due to ethical concerns: a descriptive cross-sectional study.

Authors: Elia N, Wager E; Tramèr MR
Published in: PLOSOne 2014; 9 (1); e85846
IF 2014: 3.234

Summary

Following on the «Boldt debacle», named according to the German anaesthetist Dr. Joachim Boldt that was found to have forgotten to seek formal ethical approval for 90 of his articles (Box 2), Editors in Chief of 18 journals that had published his papers declared their intention to retract these articles. In this cross-sectional study, we aimed to check whether a year and half later, these articles had been retracted and whether the retractions had been performed in adequacy with the published recommendations. Our study showed that 9 out of 88 articles (10%) had not been retracted at all (no publication of notice of retraction, no marked pdf) and could not be identified as requiring retraction by a reader unaware of the scandal. Only 5 had been retracted according to all the published requirements. The way other articles had been retracted varied regarding who took the responsibility for the retraction, marking of the pdf, or free access to the retracted article.Interestingly, 14 articles still required payment to be accessed, even thought they had been retracted!

Responsibility for both the lack of retraction and inadequate retractions were shared by editors and publishers and by the lack of a person responsible to check that the retraction was performed. Reasons ranged from personal problems of the editor, to legal threats from Dr. Boldt’s co-authors.

This article has been commented in a renowned website dedicated to research misconduct:

RetractionWatch:

http://retractionwatch.com/2013/09/10/what-happened-to-joachim-boldts-88-papers-that-were-supposed-to-be-retracted/
Fate of Articles That Warranted Retraction Due to Ethical Concerns: A Descriptive Cross-Sectional Study

Nadia Elia¹, ², Elizabeth Wager³, Martin R. Tramèr¹

¹ Division of Anaesthesiology, Geneva University Hospitals, and Medical Faculty, University of Geneva, Geneva, Switzerland, ² Institute of Social and Preventive Medicine, Medical Faculty, Geneva, Switzerland, ³ Sideview, Princes Risborough, United Kingdom

Abstract

Objective: To study journals’ responses to a request from the State Medical Association of Rheinland-Pfalz, Germany, to retract 88 articles due to ethical concerns, and to check whether the resulting retractions followed published guidelines.

Design: Descriptive cross-sectional study.

Population: 88 articles (18 journals) by the anaesthesiologist Dr. Boldt, that warranted retraction.

Method: According to the recommendations of the Committee on Publication Ethics, we regarded a retraction as adequate when a retraction notice was published, linked to the retracted article, identified the title and authors of the retracted article in its heading, explained the reason and who took responsibility for the retraction, and when the retracted article was freely accessible and marked using a transparent watermark that preserved original content. Two authors extracted data independently (January 2013) and contacted editors-in-chief and publishers for clarification in cases of inadequate retraction.

Results: Five articles (6%) fulfilled all criteria for adequate retraction. Nine (10%) were not retracted (no retraction notice published, full text article not marked). 79 (90%) retraction notices were published, 76 (86%) were freely accessible, but only 15 (17%) were complete. 73 (83%) full text articles were marked as retracted, of which 14 (16%) had an opaque watermark hiding parts of the original content, and 11 (13%) had all original content deleted. 59 (67%) retracted articles were freely accessible. One editor-in-chief stated personal problems as a reason for incomplete retractions, eight blamed their publishers. Two publishers cited legal threats from Dr. Boldt’s co-authors which prevented them from retracting articles.

Conclusion: Guidelines for retracting articles are incompletely followed. The role of publishers in the retraction process needs to be clarified and standards are needed on marking retracted articles. It remains unclear who should check that retractions are done properly. Legal safeguards are required to allow retraction of articles against the wishes of authors.

Citation: Elia N, Wager E, Tramèr MR (2014) Fate of Articles That Warranted Retraction Due to Ethical Concerns: A Descriptive Cross-Sectional Study. PLoS ONE 9(1): e85846. doi:10.1371/journal.pone.0085846

Introduction

On the 25th of February 2011, the State Medical Association of Rheinland-Pfalz, Germany, informed all affected medical journals of the results of its evaluation regarding the status of Institutional Review Board approval for research conducted by the anaesthetist Dr. Joachim Boldt. The evaluation revealed that 88 original articles authored by Dr. Boldt, and published in 18 peer-reviewed journals, lacked formal ethical approval. As a consequence, the editors-in-chief of these 18 journals signed a common statement declaring their intention to retract these articles from their journals, and to publish formal retraction notices [1].

Authoritative bodies such as the Committee On Publication Ethics (COPE) [2], and the National Library of Medicine (NLM) [3] have produced guidelines concerning the retraction of fraudulent research papers. According to COPE, retraction notices should be published and linked to the retracted article, be clearly identified as a retraction (not as a correction or a comment), identify the retracted article by including the title and authors in the retraction heading, explain who is retracting and the reason for retraction, and finally, be freely accessible. As a consequence, the editors-in-chief of these 18 journals signed a common statement declaring their intention to retract these articles from their journals, and to publish formal retraction notices [1].
retracted article as well as in the online version, be listed in the Table of Contents page." NLM does not remove the original reference of a retracted article, but updates the citation to indicate it has been retracted and adds a link to the retraction statement. The original article should be retained unchanged, except for a watermark on the PDF indicating on each page that it is “retracted”. Sox and Rennie have further recommended that, as with retraction notices, journals should provide free access to the full text of retracted fraudulent articles [5].

In earlier cases, some occurring before these recommendations were published, journals were found not to have retracted articles appropriately [5,6]. Although the importance of appropriately retracting fraudulent articles has often been underlined in Commentaries and Editorials [7–9], studies attempting to quantify retracting fraudulent articles has often been underlined in Commentaries and Editorials [7–9], studies attempting to quantify the problem are still lacking. We therefore set out to study the fate of the 88 articles by Dr. Joachim Boldt that were meant to be retracted in early summer 2011 because of lack of ethical approval.

Methods
This study is reported according to the STROBE statement for reporting cross-sectional studies.

Ethics statement
Ethical approval was not required for this study.

Study design and setting
This was a descriptive cross-sectional study. All searches and data extraction were done in January 2013.

Study selection
We selected all 88 articles listed in the document “Editors-in-Chief Statement Regarding Published Clinical Trials Conducted without IRB Approval by Joachim Boldt” (published in February 2011) for which the State Medical Association of Rheinland-Pfalz, Germany, was unable to verify approval by a competent ethics committee and had therefore recommended that they should be retracted [1].

Variables, data extraction and data sources
For each article, we checked whether all criteria for adequate retraction, as stipulated by COPE, were fulfilled [2]. In addition (following other recommendations) we checked whether the full text article was freely accessible [5], and whether the original content was preserved [10].

For each title, two authors (NE, MRT) independently recorded whether or not a retraction notice had been published in the respective journal (yes, no), whether the retraction notice was linked to the retracted article (yes, no), was listed in the journal’s table of contents (completely: complete reference of retracted article is listed in table of contents; incompletely: retraction is listed in table of contents but reference is lacking; none: retraction notice is not listed), and whether the heading of the retraction notice included the title (yes, no) and the authors (yes, no) of the retracted article.

For each retraction notice that could be retrieved, we checked whether or not the notice provided explanations concerning the reason for retraction (yes, no), and described who was responsible for the retraction (editor-in-chief, editorial board, publisher, etc).

We checked on the article PDFs whether, and how, they were marked, for instance, using a watermark indicating “retracted” across all pages. We classified the watermarks as transparent (i.e., underlying text, figures or tables were readable) or opaque (i.e., underlying text, figures and tables were obscured).

In order to assess the accessibility of retraction notices and full text retracted articles, all searches were performed from a private computer without subscription to any journal. For the searches of the full text PDF, we copied the titles of retracted articles into Google®. If a link to Pubmed was provided, that link was tried first. If a link to the full text was provided in Pubmed, that link was used. If the Pubmed link led to the full text (online or PDF), the article was classified as freely accessible through Pubmed. If the Pubmed link led to a login page requiring registration and/or a fee for access to the article, the article was classified as not freely accessible through Pubmed. For all articles that were not freely accessible through Pubmed, alternative links provided through Google were searched (for instance, ScienceDirect [http://www.sciencedirect.com/], ResearchGate [http://www.researchgate.net/] or Google Scholar [http://scholar.google.ch/]). When this secondary search was successful, the article was classified as freely accessible through alternative web sources. If it was not, it was classified as not freely accessible through alternative web sources. The same procedure was repeated to assess accessibility of retraction notices.

We considered a retraction as adequate if the following criteria were fulfilled: a retraction notice was published and linked to the retracted article, was identified as a “retraction” in the table of contents of the journal, included title and authors of the retracted article in its heading, explained who took responsibility for the retraction, gave reasons for retraction, and was freely accessible, and if the PDF of the retracted article was clearly labelled as “retracted” using a transparent watermark preserving original content and was freely accessible.

Finally, we contacted the editors-in-chief of those journals that had failed to retract articles correctly, and asked them for an explanation. If feasible, we also contacted the journal publishers and asked for explanations.

Bias
Data extraction was performed by two authors (NE, and MT) independently in order to minimise the risk of extraction errors. Clear procedures were defined before performing the searches in order to guarantee reproducibility of the findings.

Study size and statistical analyses
The study sample was defined as all trials that were authored or co-authored by Dr. Joachim Boldt and that were found by an official inquiry to warrant retraction because of lack of formal ethical approval. This is a descriptive study; there was no intention to search for associations between variables or to draw statistical inferences; therefore, no sample size calculation was performed. Results are reported as frequencies and percentages.

Results
Journals and publishers
The 88 articles were published in Anesthesia and Analgesia (22 articles), British Journal of Anaesthesia (11), Journal of Cardiothoracic and Vascular Anesthesia (9), European Journal of Anaesthesiology (8), Anesthesiology (6), Anaesthesiologie Intensivmedizin Notfallmedizin Schmerztherapie (6), Canadian Journal of Anesthesia (5), Intensive Care Medicine (5), Acta Anaesthesiologica Scandinavica (3), Der Anaesthesist (2), Annals of Thoracic Surgery (2), Critical Care Medicine (2), Thoracic and Cardiovascular Surgery (2), Medical Science Monitor (2), Minerva Anestesiologica (2), Anesthesiology (1), Journal of Craniomaxillofacial Surgery (1), and Vox Sanguinis (1) (Table 1).

The 18 journals were published by nine publishers (Table 1): Lippincott Williams & Wilkins, Wiley-Blackwell, Springer, and Elsevier (three journals each), Thieme (2), Oxford University Press, Edizioni Minerva Medica and Medical Science International (1 each). One journal (European Journal of Anaesthesiology) had changed
Table 1. Summary of 88 articles that warranted retraction.

<table>
<thead>
<tr>
<th>Journal and publisher</th>
<th>Retractions</th>
<th>Retraction notice</th>
<th>Linking to article</th>
<th>Identified as such</th>
<th>In table of content</th>
<th>Heading includes</th>
<th>Notice describes</th>
<th>Free access</th>
<th>Watermark</th>
<th>Free access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acta Anaesthesiologica Scandinavica</td>
<td>3</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES (1)</td>
</tr>
<tr>
<td>Anaesthesia</td>
<td>6</td>
<td>YES (2)</td>
<td>July</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES (1)</td>
</tr>
<tr>
<td>Anesthesiologie Intensivmedizin Notfallmedizin Schmerztherapie</td>
<td>1</td>
<td>YES</td>
<td>May</td>
<td>NO</td>
<td>1</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES (1)</td>
</tr>
<tr>
<td>Anaesthesia &amp; Analgesia</td>
<td>22</td>
<td>YES</td>
<td>May</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES (1)</td>
</tr>
<tr>
<td>Oxford University Press</td>
<td>11</td>
<td>YES</td>
<td>July</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO (1)</td>
</tr>
<tr>
<td>Intensive Care Medicine</td>
<td>5</td>
<td>YES (1)</td>
<td>Sept</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO (1)</td>
</tr>
<tr>
<td>European Journal of Anesthesiology</td>
<td>7</td>
<td>YES</td>
<td>June</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO (1)</td>
</tr>
<tr>
<td>Intensive Care Medicine</td>
<td>1</td>
<td>YES</td>
<td>June</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO (1)</td>
</tr>
<tr>
<td>Intensive Care Medicine</td>
<td>5</td>
<td>YES</td>
<td>July</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO (1)</td>
</tr>
<tr>
<td>Journal of Cardiothoracic and Vascular Anesthesia</td>
<td>9</td>
<td>YES</td>
<td>Aug</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO (1)</td>
</tr>
<tr>
<td>Journal of Cranio-Maxillofacial Surgery</td>
<td>1</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Retraction of Articles for Ethical Concerns
Table 1. Cont.

<table>
<thead>
<tr>
<th>Journal and publisher</th>
<th>Full text article (PDF)</th>
<th>Watermark</th>
<th>Free access</th>
<th>Notice describes</th>
<th>Heading includes Notice</th>
<th>Full text article (PDF)</th>
<th>Reprinted Notice</th>
<th>Free access</th>
<th>Watermark</th>
<th>Present Quality</th>
<th>Pubmed Other</th>
<th>Present Quality</th>
<th>Pubmed Other</th>
<th>Quality PubMed Other</th>
<th>Other Quality</th>
<th>Quality PubMed Other</th>
<th>Other Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Science Monitor</td>
<td>Medical Science International</td>
<td>2</td>
<td>None</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minerva Anestesiologica</td>
<td>Edizioni Minerva Medica</td>
<td>1</td>
<td>1</td>
<td>May</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Thoracic and Cardiovascular Surgeon</td>
<td>Thieme</td>
<td>1</td>
<td>None</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiley-Blackwell</td>
<td>1</td>
<td>1</td>
<td>July</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vox Sanguei</td>
<td>1</td>
<td>1</td>
<td>April</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScientoMedica</td>
<td>1</td>
<td>1</td>
<td>March</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Cardiology</td>
<td>1</td>
<td>1</td>
<td>February</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Gastroenterology</td>
<td>1</td>
<td>1</td>
<td>January</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Journal of the American Board of Internal Medicine</td>
<td>1</td>
<td>1</td>
<td>December</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Surgery</td>
<td>1</td>
<td>1</td>
<td>November</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Orthopaedic Surgery</td>
<td>1</td>
<td>1</td>
<td>October</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Sports Medicine</td>
<td>1</td>
<td>1</td>
<td>September</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Cardiology</td>
<td>1</td>
<td>1</td>
<td>August</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Surgery</td>
<td>1</td>
<td>1</td>
<td>July</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Orthopaedic Surgery</td>
<td>1</td>
<td>1</td>
<td>June</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Sports Medicine</td>
<td>1</td>
<td>1</td>
<td>May</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Cardiology</td>
<td>1</td>
<td>1</td>
<td>April</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Surgery</td>
<td>1</td>
<td>1</td>
<td>March</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Orthopaedic Surgery</td>
<td>1</td>
<td>1</td>
<td>February</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Journal of Sports Medicine</td>
<td>1</td>
<td>1</td>
<td>January</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Numbers in parentheses are numbers of articles or of retraction notices. 1: Collecting notice of retraction published in a comprehensive format of retracted article was listed in the table of contents of the respective journals; in 52 (59%; nine journals) the complete reference of the retracted article was listed in the table of contents, and in 24 (27%; four journals) the retraction was listed in the table of contents but the reference was lacking. Three published retraction notices (two journals) were not listed in the table of contents.

The formats of the retraction notices were consistent within each journal, but differed between journals (Table 1): 53 notices (60%) included the title of the retracted article, and 32 (36%) included the names of the authors in their heading. Twenty notices (23%; four journals) included both the title and the authors in their heading, and 14 (16%) included neither of them in the heading, but included the reference in the text of the retraction notice.

Seventy-one retraction notices (81%) described who had taken responsibility for the retraction. Responsibilities varied across journals. In seven journals (48 retraction notices), the editor-in-chief signed the retractions, in three (ten notices), it was the editor-in-chief with the publisher, and in two (13 notices), responsibility was taken by the editorial board. In three journals (eight notices), there was no indication of who had retracted the article; for example, one stated, “the following article has been retracted...”. Reasons for the retraction were explicitly provided in the notices in 13 journals (49 retraction notices); in two journals (30 notices), the notice referred to an editorial that explained the context of the retractions.

Sixty-seven retraction notices (76%) were accessible through Pubmed, nine retraction notices (10%; one journal) were accessible through an alternative weblink (ScienceDirect) but not through Pubmed, and three (3%; two journals) were not freely accessible. Overall, only 15 retraction notices (17%; three journals) were found to fulfill all predefined criteria for an adequate retraction notice (Fig. 1).

Retraction notices. No retraction notices had been published for nine of the articles (10%; five journals). Of these five journals, three (publishing four articles) had not published any retraction notice at all, one had published retraction notices for only two of six articles, and one had published a retraction notice for only one of two articles (Table 1). Retraction notices for the remaining 79 articles were published between May 2011 and October 2011.

Each of the 79 retraction notices was linked, in Pubmed, to the retracted article; 55 (63%) notices were clearly identified as “Retractions”, one was labelled “Statement”, one “Erratum”, and 22 (originating from one journal) were not labelled at all and referred to an Editor’s Note named: “Notice of retraction”. 76 retraction notices (86%) were listed in the table of contents of the respective journals; in 52 (59%; nine journals) the complete reference of the retracted article was listed in the table of contents, and in 24 (27%; four journals) the retraction was listed in the table of contents but the reference was lacking. Three published retraction notices (two journals) were not listed in the table of contents.

Outcome data and main results

Five retraction notices (6%; published in one journal) fulfilled all predefined criteria of adequate retraction notice (Fig. 1).

Retraction notices. No retraction notices had been published for nine of the articles (10%; five journals). Of these five journals, three (publishing four articles) had not published any retraction notice at all, one had published retraction notices for only two of six articles, and one had published a retraction notice for only one of two articles (Table 1). Retraction notices for the remaining 79 articles were published between May 2011 and October 2011.

Each of the 79 retraction notices was linked, in Pubmed, to the retracted article; 55 (63%) notices were clearly identified as “Retractions”, one was labelled “Statement”, one “Erratum”, and 22 (originating from one journal) were not labelled at all and referred to an Editor’s Note named: “Notice of retraction”. 76 retraction notices (86%) were listed in the table of contents of the respective journals; in 52 (59%; nine journals) the complete reference of the retracted article was listed in the table of contents, and in 24 (27%; four journals) the retraction was listed in the table of contents but the reference was lacking. Three published retraction notices (two journals) were not listed in the table of contents.

The formats of the retraction notices were consistent within each journal, but differed between journals (Table 1): 53 notices (60%) included the title of the retracted article, and 32 (36%) included the names of the authors in their heading. Twenty notices (23%; four journals) included both the title and the authors in their heading, and 14 (16%) included neither of them in the heading, but included the reference in the text of the retraction notice.

Seventy-one retraction notices (81%) described who had taken responsibility for the retraction. Responsibilities varied across journals. In seven journals (48 retraction notices), the editor-in-chief signed the retractions, in three (ten notices), it was the editor-in-chief with the publisher, and in two (13 notices), responsibility was taken by the editorial board. In three journals (eight notices), there was no indication of who had retracted the article; for example, one stated, “the following article has been retracted...”. Reasons for the retraction were explicitly provided in the notices in 13 journals (49 retraction notices); in two journals (30 notices), the notice referred to an editorial that explained the context of the retractions.

Sixty-seven retraction notices (76%) were accessible through Pubmed, nine retraction notices (10%; one journal) were accessible through an alternative weblink (ScienceDirect) but not through Pubmed, and three (3%; two journals) were not freely accessible. Overall, only 15 retraction notices (17%; three journals) were found to fulfill all predefined criteria for an adequate retraction notice (Fig. 1).

Retracted articles. Fifteen articles (17%) were not marked as retracted. Nine of these also lacked a retraction notice. Of the 73 articles (83%) that were marked as retracted, 48 (55%; eight journals) had transparent watermarks although in ten of those (11%; two journals), the watermark was almost invisible (Fig. 2). The other 14 articles (16%; four journals) had opaque marks across all pages that completely obscured parts of text, tables or figures (Fig. 3). Eleven articles (13%; all from one journal) had their entire content (i.e. text, tables, figures, references) deleted.

In one journal, only seven of eight full texts were labelled with a watermark although retraction notices were published for all eight.
Retraction of Articles for Ethical Concerns

Figure 1. Flow chart. 1 Retraction notice linked to the retracted article, identified as a retraction in the table of content of the journal, included the title and authors of the retracted article in its heading, explained who took the responsibility for retraction, and provided reasons for retraction. 2 Watermark transparent, not opaque, and the original text preserved. 3 Through PubMed or alternative websites. doi:10.1371/journal.pone.0085846.g001

Figure 2. Transparent watermarks. Left panel: Anesthesia and Analgesia. Right panel: Canadian Journal of Anesthesia. doi:10.1371/journal.pone.0085846.g002

TABLE II. Mean arterial blood pressure (MAP), heart rate (HR) and rectal temperature (RT)

<table>
<thead>
<tr>
<th>Group</th>
<th>T0</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unretraced</td>
<td>94 ± 13</td>
<td>98 ± 16*</td>
<td>97 ± 12</td>
<td>95 ± 12</td>
<td>91 ± 8*</td>
</tr>
<tr>
<td>MAP (mmHg)</td>
<td>36.4 ± 0.3</td>
<td>35.3 ± 0.5</td>
<td>35.7 ± 0.4</td>
<td>not done</td>
<td>35.9 ± 0.1</td>
</tr>
<tr>
<td>HR (beats/min)</td>
<td>79 ± 13</td>
<td>89 ± 12</td>
<td>84 ± 12</td>
<td>78 ± 13</td>
<td>74 ± 13</td>
</tr>
<tr>
<td>RT (C)</td>
<td>36.3 ± 0.4</td>
<td>35.4 ± 0.5</td>
<td>35.7 ± 0.4</td>
<td>not done</td>
<td>36.0 ± 0.4</td>
</tr>
<tr>
<td>Unretraced</td>
<td>93 ± 13</td>
<td>95 ± 19</td>
<td>91 ± 13</td>
<td>88 ± 19</td>
<td>88 ± 10</td>
</tr>
<tr>
<td>MAP (mmHg)</td>
<td>36.4 ± 0.3</td>
<td>35.3 ± 0.5</td>
<td>35.7 ± 0.4</td>
<td>not done</td>
<td>36.0 ± 0.4</td>
</tr>
<tr>
<td>HR (beats/min)</td>
<td>79 ± 13</td>
<td>89 ± 12</td>
<td>84 ± 12</td>
<td>78 ± 13</td>
<td>74 ± 13</td>
</tr>
<tr>
<td>RT (C)</td>
<td>36.3 ± 0.4</td>
<td>35.4 ± 0.5</td>
<td>35.7 ± 0.4</td>
<td>not done</td>
<td>36.0 ± 0.4</td>
</tr>
</tbody>
</table>

Table 2. Demographic Data and CPB Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y)</td>
<td>50.9 ± 11.8</td>
<td>17.8</td>
</tr>
<tr>
<td>Gender (male/female)</td>
<td>51/29</td>
<td>0.17</td>
</tr>
<tr>
<td>Duration of surgery (min)</td>
<td>271.5 ± 55.3</td>
<td>205.5</td>
</tr>
<tr>
<td>Temperature (-O°C)</td>
<td>34.8 ± 0.8</td>
<td>34.9 ± 0.7</td>
</tr>
<tr>
<td>Duration of CPB (min)</td>
<td>83.3 ± 32.2</td>
<td>74.26</td>
</tr>
<tr>
<td>CPB dose (U)</td>
<td>27,000 (3741)</td>
<td>25,400 (4138)</td>
</tr>
<tr>
<td>Cumulative protamine dose (mg)</td>
<td>281.0 (41.3)</td>
<td>275.0 (39.4)</td>
</tr>
</tbody>
</table>

*P < 0.05 compared with control group.
The article that lacked a watermark had been published by the journal's previous publisher.

Forty-four articles (50%; six journals) were freely accessible through Pubmed, but 11 of those (13%; one journal) had their content completely deleted. Fifteen full text articles (17%) were freely accessible, although through alternative weblinks only; one of them was not marked. Overall, only 34 articles (39%) were both adequately marked and freely accessible (Fig. 1).

Explanations from editors-in-chief and publishers. We contacted the editors-in-chief of the journals that had not correctly implemented some or all retractions and asked them for explanations. Two did not respond. One referred to personal health problems that prevented him from accomplishing the task. The others referred us to their publishers. One of these editors explained that it had been the publisher's decision to delete the content of the retracted one of two retraction notices. One editor explained that it was due to internal communication problems that led to the omission of parts of the original information. A major issue was the lack of free access to both retraction notices and full texts of retracted articles; only 76 (86%) of all retraction notices and 59 (67%) of all retracted articles were freely accessible, and not all of those were freely accessible through PubMed but had to be accessed through alternative websites.

Formal retraction from the literature is the most severe sanction for a published research article [2]. Since science must be self-correcting, retraction of unreliable articles is an essential step for rectifying the scientific knowledge-base [11], although it has been shown that it is probably insufficient to inform the scientific community [12]. Retractions appear to be “unpopular” with both editors and institutions since they may shed doubt on the integrity of science, and on the expertise of the editorial team. However, they demonstrate the determination to maintain the integrity of knowledge and to prevent readers from being misled by unreliable information [10,13]. What is certain is that rejections are, and will remain, necessary since the safeguards of science including the process of peer review, remain vulnerable to fraud and error. So why did we find such great disparity in the way in which these 88 articles were handled? There may be several explanations.

First, nine (10%) articles have not been retracted at all. There is little agreement on what exactly requires a retraction [8], and lack of ethical approval does not fit into any of the conventional definitions of misconduct such as plagiarism, fabrication, or falsification of data. Therefore, some editors and publishers may be inclined to consider ethical concerns as a minor problem only. NLM advises that articles may be withdrawn because of “pervasive error” or “unsubstantiated or irreproducible data” due to either misconduct or honest error. COPE advises journal editors to retract an article if they have clear evidence that the findings are “unreliable”, and that they should consider retracting a publication if it reports “unethical research”. The 88 articles by Dr. Boldt and co-workers do not clearly fit into any of these categories and there is not yet an agreement on how to handle a clinical research article that may not necessarily be “unethical” per se, but that was not officially approved by a competent ethics committee. It is therefore possible that retraction was delayed or prevented because the articles did not fit into one of the “usual” categories requiring retraction. However, this reason was not mentioned by any of the editors who were contacted requesting an explanation and this does not explain why the retractions failed to adhere to published guidelines.

Second, no reliable mechanisms exist to ensure that research articles warranting retraction (e.g. following an appropriate investigation) are actually retracted [7,10,14] although the COPE guidelines on cooperation between journals and institutions give
some guidance on this [4]. For example, the primary responsibility for investigating possible scientific misconduct rests with the authors’ institution. Once an institution has determined that misconduct involving a research publication has occurred, journals are obliged to consider retraction of the work [11]. In the present example, suspicion of scientific misconduct was first raised by journal editors, who then turned to the authors’ institution and its ethics committee, asking for an internal investigation. When misconduct was recognised by the competent ethics committee, the editors-in-chief of all journals involved decided to start a joint retraction process. However, the individual retraction processes were handled independently by each journal, and did not follow a clearly defined common procedure.

Third, it remains unclear who should be responsible for retracting an article. In cases of unintentional “honest” error, the responsibility for requiring a retraction rests with the authors. In cases of scientific misconduct, obviously, someone else has to assume this responsibility. It has been claimed that, once an institution has identified fraud or significant error, it is then up to the journals to respond promptly and properly [4,14]. However, there is no clear guidance about who, within the “journal”, should take this responsibility. Surveys of retractions have shown that this varies and may be the editor-in-chief, the entire editorial board, the publisher, or the owner of the journal, which may be an academic society [6]. Interestingly, retraction notices sometimes specified that an editor-in-chief alone had ordered the retraction, and sometimes it seemed to be an agreement between a publisher and an editor-in-chief or the entire editorial board. Most editors of journals that had failed to correctly retract some or all articles, referred us to their publishers, suggesting that they held them responsible. As long ago as 1990 it was recognised that authors, editors, reviewers, and librarians all needed to be involved in a multifaceted approach to address the continued use of invalid data [15]; interestingly, publishers were ignored at that time. By 2012, adequately dealing with scientific misconduct has become, according to different guidelines, a joint mission for “authors, editors, and publishers” [16–18]. The exact responsibility of each actor, however, remains ill defined.

Fourth, recommendations on how full texts of retracted articles should be labelled are also lacking. COPE, for instance, states that retracted articles should not be removed from printed copies of the journal or from electronic archives but their retracted status should be indicated as clearly as possible. The question remains, whether the original content of these articles should be preserved. Some may agree with one of the interviewed editors who argued that the data were false and therefore valueless, so why leave it viewable? An alternative argument may be that scientific data, even when fraudulent or unethical, belongs to the public and may serve future research, for instance, research into fraud, and that the preservation of the historical record, including all faults, mistakes, and corrections, is essential [10]. Indeed, in 1984, when the NLM implemented a policy for identifying and indexing published retractions, they chose to link the notice of retraction to the original article rather than delete the citation to the retracted article, because they felt that removal might affect historical perspective [10]. In the present study, data from 25 retracted articles (28% of retracted articles) had been partially or completely removed; either the contents of the articles were completely deleted, or parts of the underlying text, tables or figures were hidden by opaque watermarks.

Fifth, a further unresolved issue refers to the ultimate control after a retraction has been initiated. Who ensures that an article is adequately retracted, that a notice is published in the journal and indexed in databases such as Pubmed, that the full text article is clearly marked and freely accessible electronically, and that the original information remains visible? When studying various recommendations on how to retract articles [2], it is striking how much responsibility is given to the editors. Several guidelines indicate that editors are responsible for the final decision about retracting material, with or without cooperation of the authors, and additionally that they should ensure that retractions are labelled in such a way that they are identified by bibliographic databases. There seems to be some contradiction here since we found that some editors considered they were powerless against a publisher’s decision. Also, current guidelines do not specify who should verify whether the retraction process has been implemented correctly.

Finally, journal editors or publishers may be reluctant to issue retractions and to mark an article because they may fear legal actions by discredited authors. In our example, it was sometimes a publisher that decided not to retract an article because Boldt’s co-authors threatened legal action. This highlights, again, the central role of publishers in deciding whether an article is to be retracted or not. Today, the threat of legal action weighs heavily, especially on smaller journals [9], although, according to COPE, authors usually would not have grounds for taking legal action against a journal over the act of retraction if it follows a suitable investigation and proper procedures [2]. COPE also states that journal editors should consider at least issuing an expression of concern if an investigation is underway but a judgment will not be available for a considerable time [2]. None of the journals that failed to retract an article has published such an expression of concern.

Our analysis has some limitations. The main weakness of our study is that it remains descriptive and relates to a single series of papers involving one author and a single investigation into a specific case of lack of evidence of ethical approval, which may not be typical of most retractions; we have not attempted to identify potential “risk factors” for problems in the retraction process. The reason for this is that we had no strong a priori hypothesis to test, and the relatively small sample size would have prevented us from performing multivariate analyses. However, our analyses may serve as a basis for future larger studies focusing on retractions due to ethical issues and for other reasons. Indeed, this descriptive cross-sectional study is the first of its kind to describe systematically the disparity of the retraction processes in a uniform context; one common author for each article and a single type of misconduct (lack of ethical approval) was involved. Selection bias is unlikely since we included all articles by Dr. Boldt that had been identified by an official investigation [1]. It cannot be excluded that a few additional studies will eventually require retraction because of lack of ethical approval but it is unlikely that this will change the overall picture. The risk of reporting bias was minimized by having two researchers extract the relevant data separately.

The Boldt case was a shock to the academic world [19]. The impact of this debacle was recognized even outside peri-operative medicine [20]. Perhaps the only positive and encouraging fact was that the editors-in-chief of 18 journals agreed, in a well-orchestrated, committed and overt way, to sign a strong public statement and to retract 88 articles. This organised approach against fraud, across so many journals, was probably unique in the scientific world until then, and it has been cited as a laudable example of how journal editors should play a more active role in the retraction of fraudulent papers [8]. Since, an even larger case of scientific misconduct, necessitating the retraction of 183 articles, has been uncovered [21]. Our study shows that purging the literature of fraudulent or unethical articles remains a technically challenging process and highlights several weaknesses which must
be addressed. This is probably more a confirmation than a revelation. We did know that retractions were imperfect, but our study helps to quantify this. Perhaps most fundamentally, there must be clear and universally accepted definitions about what type of articles deserve retraction. We feel strongly that science that has not been approved by a competent ethics committee should fall into this category and we suggest that COPE's wording about "unethical research" should be amended to include this. The roles and responsibilities of the different players must be unambiguously defined, and publishers must not be left out—indeed, they should probably have a central role. The mechanism for retracting articles published by a previous publisher (i.e. after a journal has switched publishers) needs to be resolved. It seems obvious that once a competent independent investigation has provided convincing evidence that an article should be retracted, it is the journal's responsibility to issue a retraction notice, but it remains unclear who must take on that responsibility, the editor or the publisher. We suggest that it should be the publisher's responsibility to adequately mark the full text of the retracted article, using a transparent rather than opaque watermark on each page. Deleting the content of the fraudulent article should be outlawed.

Both retraction notices and retracted articles belong to the public domain and should not be hidden behind access barriers or pay walls. After a reasonable time period, it should be verified whether the retraction process has been successfully implemented, although, again, it remains unclear who shall take on that responsibility, but perhaps this role could be taken by the institution that carried out the investigation. And finally, legal safeguards are needed that allow journal editors to retract fraudulent or unethical papers even against the wishes of authors and/or publishers.

Acknowledgments
Special thanks go to the editors and publishers who responded to our enquiry and provided additional information.

Author Contributions
Conceived and designed the experiments: NE MRT. Performed the experiments: NE MRT. Analyzed the data: NE MRT EW. Wrote the paper: NE MRT EW. Read and approved the final manuscript: NE MRT EW.

References
5.3 Susceptibility to fraud in systematic reviews: lessons from the Reuben case.

Authors: Marret E, Elia N, Dahl JB, McQuay HJ, Meiniche S, Moore RA, Straube S, Tramèr, MR
Published in: Anesthesiology 2009; 111:1279-89
IF 2009: 5.354

Summary

In this study, our aim was to investigate and quantify the impact of the “Reuben fraud” (see Box 1) on the conclusions of systematic reviews that had included one or more article based on fabricated data in their analyses, since it had been claimed that the massive retraction of 20 articles by Scott Reuben compromised the conclusions of every systematic reviews that had included these trials. We wanted to check whether systematic reviews could be “robust” to fraud, or whether their conclusions differed widely from the conclusion they would have reached had they not considered the fraudulent studies. We re-analysed 25 systematic reviews that cited (5), considered (6) or included (14) a trial authored by Scott Reuben, with or without the (potentially) fraudulent data. Overall, we showed that carefully performed quantitative systematic reviews were mostly robust against the Reuben fraud, even though qualitative systematic reviews including fraudulent reports were more vulnerable regarding their conclusions, especially if the proportion of fabricated data among all included trials was higher than 30%. This study highlighted yet other interesting issues such as the unclear role of co-authors of Scott Reuben (one co-author contributed up to 8 articles, but was not bothered by the scandal), or the unclear role of sponsors (all reports that had been sponsored by the pharmaceutical industry were withdrawn). It also highlighted the high rate of self-citations in Scott Reuben’s articles. Our conclusion was that one of the strength of a systematic review was to shift emphasis from single studies to multiple ones, and biased conclusion could be avoided if some basic principles of systematic reviewing were observed.
Susceptibility to Fraud in Systematic Reviews

Lessons from the Reuben Case


Background: Dr. Scott Reuben allegedly fabricated data. The authors of the current article examined the impact of Reuben reports on conclusions of systematic reviews.

Methods: The authors searched in ISI Web of Knowledge systematic reviews citing Reuben reports. Systematic reviews were grouped into one of three categories: I, only cited but did not include Reuben reports; II, retrieved and considered, but eventually excluded Reuben reports; III, included Reuben reports. For quantitative systematic reviews, each author decided independently whether noninclusion of Reuben reports would have changed conclusions.

Results: Twenty-five systematic reviews (5 category I, 6 category II, 14 category III) cited 27 Reuben reports (published 1994–2007). Most tested analgesics in surgical patients. One of 6 quantitative category III reviews would have reached different conclusions without Reuben reports. In all 6 (30 subgroup analyses involving Reuben reports), exclusion of Reuben reports never made any difference when the number of patients from Reuben reports was less than 30% of all patients included in the analysis. Of 8 qualitative category III reviews, all authors agreed that one would certainly have reached different conclusions without Reuben reports. For another 4, the authors' judgment was not unanimous.

Conclusions: Carefully performed systematic reviews proved robust against the impact of Reuben reports. Quantitative systematic reviews were vulnerable if the fraudulent data were more than 30% of the total. Qualitative systematic reviews seemed at greater risk than quantitative.

SYSTEMATIC reviews of randomized controlled trials (RCTs), with or without meta-analysis, are considered powerful levels of evidence on which to guide clinical practice. However, bias and fraud in original research can threaten systematic review and meta-analysis. For example, when systematic reviews include trials with inadequate concealment of treatment allocation or inappropriate blinding, they are likely to overestimate the benefit of a treatment.1,2 Similarly, accidental inclusion of covert duplicate publication into meta-analysis can bias the conclusion of that meta-analysis in favor of an experimental intervention.3

Falsification of data is a grave breach of scientific ethics. Fabricated data may be published in peer-reviewed scientific journals and may subsequently be included in systematic reviews. Recently, routine audit uncovered perhaps one of the largest research frauds ever reported.4 U.S. anesthesiologist Dr. Scott Reuben allegedly fabricated clinical studies.5 Most of these trials demonstrated benefits from analgesic drugs.

Data from Reuben publications have been included in systematic reviews and meta-analyses. It has been claimed that the retraction of Reuben’s articles compromised every systematic review that included these fabricated findings.6 In this context, two questions are relevant. First, does noninclusion of fraudulent data in systematic reviews change the conclusions of these systematic reviews? Second, are some systematic reviews more robust than others against the impact of included fraudulent data, and if so why? We set out to address these issues using the Reuben case as an example.

Materials and Methods

For the purpose of this analysis, we regarded all reports that were coauthored by Reuben as potentially fraudulent, including those not included on the official retraction list.7

We searched for indexed reports of any study architecture that were coauthored by Dr. Scott S. Reuben. We searched ISI Web of Knowledge using the author search...
term Reuben SS.‡‡ The date of the last search was March 18, 2009. The Create Citation Report tool was used to summarize bibliometric data of the Reuben reports. From all reports that cited Reuben at least once, we selected those that used the term systematic review or meta-analysis in the title. When the title left doubt about the nature of the citing reference, we consulted the abstract.

We checked whether the citing reviews fulfilled at least one of two criteria of a systematic review: (1) a methods section with a description of the search strategy or (2) explicit inclusion criteria for eligible reports. All other reviews were regarded as narrative, nonsystematic reviews and were not considered further.

Systematic reviews were categorized into three subgroups: Category I cited a Reuben report (e.g., in the introduction or the discussion) but did not consider the data for inclusion (n=5). Category II retrieved and considered a Reuben report for inclusion but eventually excluded it on the basis of, for instance, quality or validity criteria (n=6). Category III included data from a Reuben report either quantitatively (i.e., meta-analytically) or qualitatively.

**Results**

We retrieved 96 Reuben reports. Sixty-four were cited at least once (total number of citations, 1,199; without self-citations, 682 or 57% of all citations). Of the 64 cited Reuben reports, 27 (see Supplemental Digital Content 1, a document that lists all Reuben references cited in this study, http://links.lww.com/ALN/A559) were cited by at least one of 28 reviews (fig. 1). There was a pre boc agreement that a change from a significant to a nonsignificant result (or vice versa) by excluding Reuben data were a relevant change in the outcome of a meta-analysis.

**Fig. 1. Systematic reviews that cited Reuben reports.** No search strategy described.

---

patients per trial, 60 [range, 40–200]). One was a retrospective analysis including data from 434 patients, and 1 was performed in 15 healthy volunteers. Ten reports were officially withdrawn. Seven of the 10 officially withdrawn reports tested a coxib. Types of surgery were mainly orthopedic. Reuben was the first author of 21 reports (78%). Thirty-eight individuals acted as coauthors (median number per report, 2 [range, 1–5]). One individual coauthored eight times, one coauthored six times, and one coauthored five times.

Category I Systematic Reviews

Five systematic reviews (studying antiepileptic drugs against morphine-induced emesis, perioperative gabapentin and pregabalin, or postdischarge symptoms after outpatient surgery) retrieved and considered Reuben reports. This was mainly due to a decrease in power, i.e., 95% confidence intervals became wider. However, for some subgroup analyses, point estimates also changed. For example, 24-h morphine sparing with coxibs changed from an average of −30.2 mg to −2 mg after exclusion of Reuben reports. Similarly, perioperative blood loss with NSAIDs or celecoxib changed from −19.7 ml to +23.7 ml (table 2).

In the 6 meta-analyses, 30 subgroup analyses involved Reuben reports. In 19 (65%), the ratio of the numbers of Reuben reports over the numbers of all trials and the ratio of the numbers of patients in Reuben reports over the numbers of all trials were approximately 30% or lower (fig. 2); for none did the exclusion of Reuben reports make any difference. In 8 analyses (27%), the ratios of the numbers of reports and patients were between approximately 40% and 70%, and for 5 of those, exclusion of Reuben reports made a relevant difference. In 3 analyses (10%), the ratios of the numbers of reports and patients were 80% and higher, and for all 3, exclusion of Reuben reports made a relevant difference. All analyses with significant changes in results after exclusion of Reuben reports were from 1 single meta-analysis.

Qualitative Systematic Reviews. It was our unanimous verdict for 3 of 8 qualitative systematic reviews that noninclusion of Reuben reports would not have changed their conclusion. They studied the role of clonidine as an adjuvant to local anesthetics for periph-
eral nerve blockade, adverse effects associated with opioids, and controlled-release oxycodone for the treatment of cancer and noncancer pain. It was our unanimous verdict that noninclusion of Reuben reports would have had an impact on the results of 1 qualitative systematic review. This analysis of adjuvants to local anesthetics for intravenous regional anesthesia (IVRA) included 29 trials (1,217 patients); 6 (325 patients) were from Reuben [4,6,9,11,12,15]. The authors pointed out that only 1 trial (a Reuben report [4]) looked at the potential intraoperative benefit of NSAIDs added to local anesthetics: significantly fewer patients had tourniquet pain when ketorolac was added, and reportedly there were a number of significant post-

<table>
<thead>
<tr>
<th>Authors, Journal, Year</th>
<th>Reference No.</th>
<th>Architecture of Report</th>
<th>No. of Patients in Report</th>
<th>No. of Citations</th>
<th>No. of Coauthors</th>
<th>No. Citing Systematic Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuben SS, Duprat KM. Reg Anesth 1996</td>
<td>[6]</td>
<td>RCT</td>
<td>60</td>
<td>16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Reuben SS, Ekman EF. Anesth Analg 2007</td>
<td>[26]</td>
<td>RCT</td>
<td>191</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Reuben SS, Ekman EF, Charron D. Anesth Analg 2007</td>
<td>[27]</td>
<td>RCT</td>
<td>200</td>
<td>9</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
operative benefits from IVRA with ketorolac compared with systemic control. According to these authors, another Reuben report [6] found that ketorolac was equally analgesic either infiltrated into the surgical site or when given as an adjunct to IVRA. In addition, they thought, based on a Reuben report that was performed in 15 volunteers [15], that clonidine prolonged tourniquet tolerance and improved postoperative analgesia. According to the authors, these experimental findings were further supported by a clinical study by Reuben [12]. They also stressed that according to various Reuben reports [12,15], a small dose of clonidine as an adjuvant to IVRA seemed to be well tolerated. Finally, they concluded that opioids were disappointing for

Table 1. Continued

<table>
<thead>
<tr>
<th>No. of Patients per Group</th>
<th>Withdrawn</th>
<th>Sponsorship</th>
<th>Experimental Intervention</th>
<th>Route of Administration</th>
<th>Setting, Type of Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 No Not specified</td>
<td>Ketorolac</td>
<td>Intrathecal</td>
<td>Lower extremity vascular</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 No Not specified</td>
<td>Ketorolac</td>
<td>IVRA</td>
<td>Hand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 No Not specified</td>
<td>Ketorolac</td>
<td>Intraarticular</td>
<td>Knee arthroscopy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Yes Not specified</td>
<td>Ketorolac</td>
<td>Intraarticular</td>
<td>Knee arthroscopy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 No Not specified</td>
<td>Ketorolac</td>
<td>IVRA, wound infiltration</td>
<td>Hand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 No Not specified</td>
<td>Ketorolac</td>
<td>Morphine PCA</td>
<td>Spine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 No Not specified</td>
<td>Ketorolac</td>
<td>Intravenous</td>
<td>Spinal fusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 No Not specified</td>
<td>Ketorolac</td>
<td>IVRA</td>
<td>Carpal tunnel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Yes Not specified</td>
<td>Oxycodone</td>
<td>Oral</td>
<td>Anterior cruciate ligament reconstruction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 No Institutional and/or departmental sources</td>
<td>Clonidine</td>
<td>IVRA</td>
<td>Carpal tunnel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 No Not specified</td>
<td>Meperidine</td>
<td>IVRA</td>
<td>Carpal tunnel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 No Not specified</td>
<td>Clonidine</td>
<td>Adjuvant to local anesthetic</td>
<td>Eye</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 No Institutional and/or departmental sources</td>
<td>Clonidine, rofecoxib</td>
<td>Oral</td>
<td>Spinal fusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 No Not specified</td>
<td>Clonidine, morphine</td>
<td>Intraarticular</td>
<td>Knee arthroscopy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 No Not specified</td>
<td>Bupivacaine, morphine</td>
<td>Intraarticular</td>
<td>Knee arthroscopy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 No Not specified</td>
<td>Rofecoxib</td>
<td>Oral</td>
<td>Knee arthroscopy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 Yes Merck</td>
<td>Rofecoxib</td>
<td>Oral</td>
<td>Total knee arthroplasty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32, 34 No Not specified</td>
<td>Rofecoxib</td>
<td>Oral</td>
<td>Pediatric tonsillectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47, 48 Yes Rays of Hope Foundation, Wyeth-Ayerst Laboratories</td>
<td>Venlafaxine</td>
<td>Oral</td>
<td>Mastectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 Yes Pfizer</td>
<td>Celecoxib, Ketorolact, celecoxib, rofecoxib</td>
<td>Oral</td>
<td>Spinal fusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>130, 124 No Institutional and/or departmental sources</td>
<td>Celecoxib, Ketorolact, celecoxib, rofecoxib</td>
<td>Oral</td>
<td>Spinal fusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60, 120 Yes Pfizer</td>
<td>Celecoxib</td>
<td>Oral</td>
<td>Spinal fusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Yes Not specified</td>
<td>Celecoxib, pregabalin</td>
<td>Oral</td>
<td>Spinal fusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA Yes Pfizer</td>
<td>Celecoxib</td>
<td>Oral</td>
<td>Anterior cruciate ligament reconstruction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 Yes Pfizer</td>
<td>Celecoxib</td>
<td>Oral</td>
<td>Anterior cruciate ligament reconstruction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For references of Reuben reports [shown in brackets], please refer to Supplemental Digital Content 1, http://links.lww.com/ALN/A559.

IVRA = intravenous regional anesthesia; NA = not available; PCA = patient-controlled analgesia; RCT = randomized controlled trial.
IVRA and that, based on data from Reuben [11], only meperidine had substantial postoperative benefit but at the expense of postdeflation side effects. It was our view that noninclusion of the 6 Reuben reports would have substantially changed some of the conclusions of that qualitative systematic review.

For 4 qualitative category III systematic reviews, our verdict of whether noninclusion of Reuben reports...
would have had an impact on the results or the conclusions of the review was not unanimous.15,20,27,30 These are discussed briefly, and the reasons for our lack of unanimity are explained.

Rømsing and Møiniche27 compared four different coxibs with NSAIDs for postoperative analgesia. They included 3 Reuben reports [16,18,20] and excluded an additional report [19] because pain intensity up to 24 h
Reuben reports. All squares significant changes in results after exclusion of Reuben reports; meta-analysis. Exclusion of Reuben reports never had a significant impact on results when ratios of numbers of reports and of patients were less than 30%.

Anesthesiology, V 111, No 6, Dec 2009

Fig. 2. Ratio of numbers of reports and ratio of numbers of patients. Ratio of number of Reuben reports over number of all reports in a meta-analysis (x-axis) and ratio of number of patients in Reuben reports over number of all patients of all reports in a meta-analysis (y-axis). Data are from six meta-analyses (30 subgroup analyses) that included Reuben reports. Each symbol represents one subgroup analysis that included data from at least one Reuben report. Light gray circles = subgroup analyses with significant changes in results after exclusion of Reuben reports; dark squares = subgroup analyses that were unaffected by exclusion of Reuben reports. All light gray circles (n = 8) are from one single meta-analysis.19 Exclusion of Reuben reports never had a significant impact on results when ratios of numbers of reports and of patients were less than 30%.

postoperatively was not reported. Our ambiguity was due to one of the main conclusions of the review, stating that 50 mg rofecoxib provided superior analgesia compared with 200 mg celecoxib. That conclusion was based on data from 4 trials; 2 were from the same group of Merck collaborators, and 1 was from Reuben [16], and all 3 were in favor of rofecoxib.

Straube et al.20 tested the effect of coxibs on postoperative outcomes. They included 4 Reuben reports that tested celecoxib or rofecoxib [16,18–20]. The authors used a vote counting procedure. Noninclusion of Reuben reports would not have changed the ratio of positive to negative trials. However, in the discussion the authors stated that “With one exception (citing a Reuben report [18]) studies did not address the question of preemptive analgesia, where preoperative coxib was compared with postoperative coxib, though that exception found a large benefit for preoperative over postoperative rofecoxib,” and “Only one small study (citing a Reuben report [18]) evaluated preoperative with postoperative coxib in a standard preemptive design, and did find significant benefit for preoperative use.” Our disagreement was whether without referring to that Reuben report the authors would have been able to make a statement in favor of preemptive analgesia with rofecoxib.

Liu and Wu20 provided a summary of postoperative analgesia practices using data from systematic reviews but also from some additional, selected RCTs. The authors described a search strategy; however, selection criteria remained obscure and the reader was left in the dark as to how the conclusions from, e.g., a systematic review were weighted compared with data from a single RCT. In the results section, the authors suggested that the use of coxibs might result in a reduction in long-term complications after surgery including chronic pain. As evidence for this assumption, 2 RCTs from Reuben [26,27] were cited that both studied the analgesic efficacy of celecoxib. Although the authors did not retain this hypothesis in the conclusions of the review, our disagreement was whether these Reuben reports would have had an impact on one of the results of that review.

Finally, Fischer et al.15 published evidence-based recommendations for analgesia techniques after total knee arthroplasty. The authors acknowledged support by an educational grant from Pfizer, reimbursement by Pfizer for attending working group meetings, help and expertise in performing literature searches by a Pfizer employee, and editorial assistance by medical writers who were sponsored by Pfizer. The search strategy considered exclusively RCTs. Two Reuben reports were retrieved but were eventually excluded by the authors because surgery was not knee arthroplasty [10], or postoperative pain scores were not reported [19]. Because no Reuben reports were considered in the actual analysis, we considered whether to classify the review as category II. However, in the recommendations section, the authors unexpectedly referred to 2 further, previously not considered Reuben reports [22,23]. One was a retrospective chart review [23]. Based on those 2 reports, the authors weighted the evidence regarding bone healing against NSAIDs and in favor of coxibs: “Limited data show that conventional NSAID may have dose- and duration-dependent detrimental effects on bone healing,” and “Although there is concern about impairment of bone healing with cyclooxygenase 2-selective inhibitors, limited evidence shows that they have no detrimental effects.” Both Reuben reports were classified as level 1 evidence by the authors, and because both reports were performed in spinal surgery, the evidence was regarded as “transferable.” Our disagreement was whether noninclusion of these Reuben reports would have changed one of the main conclusions of that review.

Discussion

The majority of quantitative systematic reviews (meta-analyses) proved to be robust against the impact of potentially fraudulent Reuben reports. This was not an unexpected result because data from a few trials, even when flawed or fabricated, should have no substantial impact on the conclusions of a meta-analysis, which often includes data from hundreds or thousands of patients from a large number of trials. However, some systematic reviews seemed to be less robust; they would probably have reported on different results or would have drawn different conclusions had they not included Reuben
There were three main reasons for this lack of robustness. First, the numerical relation between the number of potentially fraudulent and the number of valid data in a systematic review seems to be crucial (although we were only able to show this empirically for meta-analyses). It may be inferred from figure 2 that meta-analyses that include mainly trials or patient data from one or only a few authors or centers need to be interpreted cautiously. Whether the cutoff ratio of approximately 30% is universally applicable depends on several factors. For example, here, trial sizes were very similar. Hence, one mega-trial would overwhelm even a large number of fabricated trials if they were small. There are no rules as to how many trials a valid meta-analysis should include. The proportion of potentially fraudulent data among valid data needs to be considered.

Our lack of unanimity in estimating the impact of Reuben reports in some qualitative systematic reviews was mainly due to two reasons. First, some systematic reviewers seemed to give undue weight to evidence emerging from particular Reuben reports. This problem is inherent to the process of qualitative systematic reviews because they cannot take into account effect size. Second, in one review that claimed to consider exclusively data from RCTs, the authors referred to observational data from Reuben that eventually had a strong impact on the overall conclusions. The Improving the Quality of Reports of Meta-Analyses of Randomized Controlled Trials (QUOROM) statement stipulates that the results of a systematic review should be interpreted in the “light of the totality of available evidence.” To avoid any misinterpretation, the QUOROM statement may need to be more specific in that the results of a systematic review should be interpreted in the “light of the totality of evidence that was retrieved through the systematic literature search.” Data not directly relevant should be interpreted as hypothesis generating and may be used as a basis for a rational research agenda rather than as transferable evidence.

Of the 96 Reuben reports, one third (n = 32) have never been cited; although their impact on science and clinical practice is impossible to quantify, we may assume that it remains low. Two thirds have been cited almost 1,200 times in indexed journals, and Reuben and his coauthors have actively participated in this dissemination process because almost half of all citations were due to self-citation. Thirty-seven reports (38%) were cited in articles other than systematic reviews; among those were editorials, guidelines, clinical studies, and conventional, nonsystematic review articles. A previously published opinion statement attempted to estimate the impact of Reuben reports in this literature. Finally, 27 reports (28%) were cited in systematic reviews. Clearly, the detrimental impact of Reuben reports on systematic reviews, if there was any, was limited for several reasons. First, Reuben reports were published over a period of 15 yr with scope for others to confirm or refute the findings. There is empirical evidence that the median survival time without substantive new evidence for meta-analyses is approximately 5 yr only, and that clinically important evidence that alters conclusions about the effectiveness and harms of treatments can accumulate rapidly. Second, the Reuben studies were of limited size, minimizing their quantitative impact on meta-analyses. Third, most Reuben reports echoed current knowledge and did not contradict science.

One particularity of a few Reuben reports was that they pretended to add new insights; these were often attractive, welcomed by many, and sometimes they seemed to be revolutionary and to advance science. Among those were the absence of detrimental effects of coxibs on bone healing after spine surgery, the beneficial long-term outcome after preemptive administration of coxibs including an allegedly decreased incidence of chronic pain after surgery, and the analgesic efficacy of ketorolac or clonidine when added to local anesthetics for intravenous regional anesthesia. Clinical algorithms based on this evidence need to be revised.

Our analysis has limitations. First, we were unable to address the role of Reuben’s coauthors. It remains obscure why a small number of individuals coauthored a large number of Reuben reports without having any suspicion about the nature of the data. Although the publishing role of senior members of a collaboration group on an article has been recognized for a while, the responsibilities of coauthors, and thus each author’s contributions, have only been emerging recently. Second, we were not able either to evaluate the impact of sponsorship. Research that is sponsored by industry may draw undue conclusions in favor of the industrial product. Some Reuben reports acknowledged sponsorship by industry; however, we do not know whether industry simply funded the authors or was actively involved in design, analysis, and publication of the studies. All reports that acknowledged sponsorship from industry were retracted because the reported data were identified as having been fabricated (table 1). For the majority of Reuben reports, no information on sponsorship was provided; we do not know whether there was none or whether it was not reported. Given the concerns about financial ties in research, one would suggest that if a substantial proportion of evidence was derived from a single study sponsor, the review of the evidence should be considered with greater skepticism. Third, we assumed that all Reuben reports that were cited by systematic reviews were fabricated and that the entire data sets were flawed. We cannot exclude that some of these reports contained valid data. However, it may be argued that it is dangerous when we discover a fraudulent research article to assume that there were no problems with all previous work.
All human activity is associated with misconduct. It is perhaps naive to believe that fraud can be avoided; it will probably always exist. Almost 2% of scientists admitted to have fabricated, falsified, or modified data at least once, and this estimate was considered to be conservative. A fraudulent article looks much the same as a nonfraudulent one; there seem to be no obvious alert signs to hint that an article is fraudulent. One of the strengths of systematic review is that it leads to a shift of emphasis from single studies to multiple studies. Unless a fraudulent study is very large and reports on an important number of events, it will not have much scope to change the conclusions of a systematic review. Consequently, conclusions from a single study should not be overestimated. Also, every effort must be undertaken to further improve quality and validity of systematic reviews. Well-conducted systematic reviews allow a more objective appraisal of the evidence than traditional, non-systematic reviews, provide a more precise estimate of a treatment effect, and may explain heterogeneity between the results of individual studies beyond the play of chance. Ill-conducted systematic reviews, on the other hand, may be biased because of exclusion of relevant studies or inclusion of inadequate studies.

Our study shows clearly that meta-analysis is not an appropriate instrument to detect fraud if the fraudulent data are in line with valid data. Instances where noninclusion of Reuben reports changed the results from pooled analyses were almost always due to a decrease in the power of the analysis related to a decrease in the amount of analyzable data (table 2). Only rarely did a point estimate change significantly.

In conclusion, misleading systematic reviews can generally be avoided if a few basic principles are observed. However, the QUOROM statement, intended to improve the quality of reporting of systematic reviews, does not significantly.

In conclusion, misleading systematic reviews can generally be avoided if a few basic principles are observed. However, the QUOROM statement, intended to improve the quality of reporting of systematic reviews, does not significantly.

In conclusion, misleading systematic reviews can generally be avoided if a few basic principles are observed. However, the QUOROM statement, intended to improve the quality of reporting of systematic reviews, does not significantly.

In conclusion, misleading systematic reviews can generally be avoided if a few basic principles are observed. However, the QUOROM statement, intended to improve the quality of reporting of systematic reviews, does not significantly.

In conclusion, misleading systematic reviews can generally be avoided if a few basic principles are observed. However, the QUOROM statement, intended to improve the quality of reporting of systematic reviews, does not significantly.


38. Eisenach JC: Data fabrication and article retraction: How not to get lost in the woods. Anesthesiology 2009; 110:955–6


43. Smith R: Investigating the previous studies of a fraudulent author. BMJ 2005; 331:288–91


46. Trikalinos NA, Evangelou E, Ioannidis JPA: Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers. J Clin Epidemiol 2008; 61:464–70
5.4 How do authors of systematic reviews deal with research misconduct in original studies? A cross-sectional analysis of systematic reviews and survey of their authors.

Authors: Elia N, VonElm E, Chatagner A, Pöpping DM, Tramèr MR
Published in: BMJ Open 2016; 6: e010442
IF 2015: 2.562

Summary

In this study our aim was to explore whether systematic reviewers were aware of the potential for research malpractice in the trials included in their analyses. Additionally, we explored whether systematic reviewers sometimes suspected new cases of research misconduct during the review process. We performed a cross-sectional analysis of 118 systematic reviews published in 2013 in one of 4 major medical journals and the Cochrane Library, and performed a survey of their authors for additional questions. Our study showed that more than half of systematic reviewers searched for duplicate publications (69%), for unpublished trials (66%), or contacted study authors (62%). However only 27% were concerned with the risk of sponsor bias, 4% with the risk related to conflicts of interest of study authors and only 2.5% checked for ethical approval of included trials. Interestingly, in 7 reviews were cases of misconduct suspected, but only 2 reported it in the published review.

This article has been commented in two renowned websites dedicated to research misconduct:

RetractionWatch:
http://retractionwatch.com/2016/08/16/should-systematic-reviewers-report-suspected-misconduct/

Rédaction Médicale et Scientifique:
http://www.h2mw.eu/redactionmedicale/revue-syst%C3%A9matique-m%C3%A9ta-analyse/
How do authors of systematic reviews deal with research malpractice and misconduct in original studies? A cross-sectional analysis of systematic reviews and survey of their authors

Nadia Elia,1,2 Erik von Elm,3 Alexandra Chatagner,4 Daniel M Pöpping,5 Martin R Tramèr1,6

ABSTRACT
Objectives: To study whether systematic reviewers apply procedures to counter-balance some common forms of research malpractice such as not publishing completed research, duplicate publications, or selective reporting of outcomes, and to see whether they identify and report misconduct.

Design: Cross-sectional analysis of systematic reviews and survey of their authors.

Participants: 118 systematic reviews published in four journals (Ann Int Med, BMJ, JAMA, Lancet), and the Cochrane Library, in 2013.

Main outcomes and measures: Number (%) of reviews that applied procedures to reduce the impact of: (1) publication bias (through searching of unpublished trials), (2) selective outcome reporting (by contacting the authors of the original studies), (3) duplicate publications, (4) sponsors’ and (5) authors’ conflicts of interest, on the conclusions of the review, and (6) looked for ethical approval of the studies. Number (%) of reviewers who suspected misconduct are reported. The procedures applied were compared across journals.

Results: 80 (68%) reviewers confirmed their data. 59 (50%) reviews applied three or more procedures; 11 (9%) applied none. Unpublished trials were searched in 79 (66%) reviews. Authors of original studies were contacted in 73 (62%). Duplicate publications were searched in 81 (69%). 27 reviews (23%) reported sponsors of the included studies; 6 (5%) analysed their impact on the conclusions of the review. Five reviews (4%) looked at conflicts of interest of study authors; none of them analysed their impact. Three reviews (2.5%) looked at ethical approval of the studies. Seven reviews (6%) suspected misconduct; only 2 (2%) reported it explicitly. Procedures applied differed across the journals.

Conclusions: Only half of the systematic reviews applied three or more of the six procedures examined. Sponsors, conflicts of interest of authors and ethical approval remain overlooked. Research misconduct is sometimes identified, but rarely reported. Guidance on when, and how, to report suspected misconduct is needed.

Strengths and limitations of this study

This study combines quantitative and qualitative methods to investigate how systematic reviewers deal with research malpractice and misconduct. It proposes clear and reproducible samples of systematic reviews from four major medical journals and the Cochrane Library. The extracted data were confirmed by 70% of the authors of the systematic reviews analysed. The systematic reviewers were not asked to confirm all their data, but only the information considered ambiguous.

There is currently no common definition of ‘research misconduct’. This may have led to an underestimation of its real prevalence.

INTRODUCTION
Rationale

Research misconduct can have devastating consequences for public health1 and patient care.2,3 While a common definition of research misconduct is still lacking, there is an urgent need to come up with strategies to prevent it.4,5 Fifteen years ago, Smith6 proposed ‘a preliminary taxonomy of research misconduct’ describing 15 practices ranging from ‘minor’ to ‘major’ misconduct. Some of these practices, however, are very common and may not be regarded by all as ‘misconduct’. Therefore, this research defines ‘malpractice’ as relatively common and minor misconduct, while the term ‘misconduct’ is used for data fabrication, falsification, plagiarism or any other intentional malpractices.

It has been shown that some of the 15 malpractices described by Smith, threaten the conclusions of systematic reviews. Examples of such malpractices include: avoiding the publication of a completed research,7,8
duplicate publications, select reporting on outcomes or adverse effects, and presenting biased results that are in favor of the sponsors or the authors.

Rigorous systematic review methodology includes specific procedures that can counter-balance some of the research malpractice. Unpublished studies may, for example, be identified through exhaustive literature searches, and statistical tests or graphical displays such as funnel plots can quantify the risk of publication bias. Unreported outcomes may be unearthed by contacting the authors of original articles, and multiple publications based on the same cohort of patients can be identified and excluded from analyses. Authors of systematic reviews can also use sensitivity analyses to quantify the impact sponsors' and authors' personal interests have on the conclusions of a review. Finally, it has been suggested that, as part of the process of systematic reviewing, ethical approval of included studies or trials should be obtained in order to identify unethical research. Systematic reviewers could hence act as whistle-blowers when reporting any suspected misconduct.

Objectives
The aim of this study is to examine whether systematic reviewers apply the aforementioned procedures, and whether they uncover and report on cases of misconduct.

The study first examines whether reviewers searched for unpublished studies or tested for publication bias, contacted authors to unearth unreported outcomes, searched for duplicate publications, analysed the impact of sponsors or possible conflicts of interest of study authors, checked on ethical approval of the studies and reported on misconduct. The secondary objective was to examine whether four major journals and the Cochrane Library reported consistently on the issue.

METHODS
The reporting of this cross-sectional study follows the STROBE recommendation. The protocol is available from the authors.

Study design
We conducted a cross-sectional analysis of systematic reviews published in 2013 in four general medical journals (Annals of Internal Medicine, BMJ, JAMA, and The Lancet). A random sample of new reviews was drawn from the Cochrane Database of Systematic Reviews in 2013, as Cochrane reviews are considered the gold standard in terms of systematic reviewing.

Setting and selection of systematic reviews
Systematic reviews were identified through a PubMed search in August 2014, using the syntax ‘systematic review [Title] AND journal title [Journal], limit 01.01.2013 to 31.12.2013’. A computer-generated random sequence was used to select 25 reviews published in 2013 in the Cochrane Library.

Reviews were selected by one of the five authors (NE) on the basis of the review titles and abstracts. This was checked by another author (AC). To be eligible, reviews had to describe a literature search strategy and include at least one trial or study. Narrative reviews or meta-analyses without an exhaustive literature search were not considered.

Variables
From each systematic review, we extracted the following information: the first author's name and country of affiliation; the number of co-authors; the name of the journal; the title of the review; the number of databases searched; the number of studies and study designs included; the language limitations applied; whether or not a protocol was registered and freely accessible; and, finally, sources of funding and possible conflicts of interest of the reviewers.

Furthermore, we examined whether each of the selected reviews applied the following six procedures, they: (1) searched for unpublished trials; (2) contacted authors to identify unreported outcomes; (3) searched for duplicate publications (defined as a redundant republication of an already published study, with or without a cross-reference to the original article); (4) analysed the impact of the sponsors of the original studies on the conclusions of the review; (5) analysed the impact of possible conflicts of interest of the authors on the conclusions of the review; and (6) extracted information on ethical approval of included studies. We used the following rating system: 0=procedure not applied, 1=partially applied, 2=fully applied (table 1).

Finally, we collected information on whether the systematic reviewers suspected, and explicitly reported on, any misconduct in the included articles.

Bias
Data from the reviews were extracted by one author (NE), and copied into a specifically designed spreadsheet. Two of the co-authors checked the data (AC and DMP). We contacted all the corresponding authors of the reviews and asked them to confirm our interpretation of their methods of review. This included their method regarding the search for unpublished trials, their contacts with the authors, their search for duplicate publications and their identification of misconduct. When there was discrepancy between our interpretation and the reviewers' answers, we used the latter. This was
done by email, and a reminder was sent to those who had not replied within 2 weeks.

**Sample size**
The capacity of systematic reviewers to identify misconduct is unknown. Our hypothesis was that 5% of systematic reviewers would identify misconduct. Therefore, we needed a minimum of 110 systematic reviews to allow us to detect a prevalence of 5%, if it existed, with a margin of error of 4% assuming an α-error of 0.05.

**Statistical methods**
Descriptive results are reported as numbers (proportions) and median (IQR) as required. To check whether systematic reviews were different from one journal to the other, we performed all descriptive analyses separately according to title of the journal. \( \chi^2 \) or Kruskal-Wallis tests were applied to test the null hypothesis of homogeneous distribution of characteristics and outcomes. We compared reviews from reviewers who answered our inquiry with reviews from those who did not, and across journals. Since Cochrane reviews were expected to be different from those published in the journals, we performed separate analyses with and without Cochrane reviews. We did not expect missing data. Statistical significance was defined as an α-error of 0.05 or less in two-sided tests. Analyses were performed using STATA V.13.

**RESULTS**

**Selection of reviews**
We identified 136 references; 18 were excluded for different reasons, leaving us with 118 systematic reviews (Ann Int Med 39\(^{A1-39}\); BMJ 38\(^{31-38}\); JAMA 12\(^{11-12}\); Lancet 10\(^{1-10}\); Cochrane Library 19\(^{C1-19}\)) (figure 1, online supplementary appendix table 1A).

**Characteristics of the reviews**
The characteristics of the reviews are described in table 2, online supplementary appendix tables 1A and 2A. Approximately 75% of the first authors were affiliated to an English-speaking institution. The protocols of all the Cochrane reviews were registered and available. However, protocols were available for only 17 reviews from the journals.

Sources of funding were declared in 110 reviews. Among these 110, 24 declared that they had no funding at all. All the reviews declared presence or absence of conflicts of interest of the reviewers.

The median number of databases searched was four. Additional references were searched in systematic reviews published previously and through contacting experts and/or authors of the original studies. Forty-two (36%) reviews only considered the English literature, and 6 (5%) reviews searched in Medline only. Four (3%) reviews searched for English articles in Medline only.\( ^\text{A}(35,39,4(5,9)) \) The median number of articles included per review was 28. Half of the systematic reviews included a mix of various study designs, while 39% included only RCTs (table 2).

**Outcome data**

**Contact with reviewers**
Out of the 118 reviews, we were able to contact 111 corresponding authors. No valid email address was available for seven. Eighty reviewers (72%) responded to our inquiries.

Among the 80 reviewers who responded, 8 (10%) provided information that changed our data extraction

<table>
<thead>
<tr>
<th>Table 1 Rating of the six procedures examined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
</tr>
<tr>
<td>Search of unpublished trials and/or test for publication bias</td>
</tr>
<tr>
<td>Contact with study authors</td>
</tr>
<tr>
<td>Duplicate publications</td>
</tr>
<tr>
<td>Sponsors of the studies</td>
</tr>
<tr>
<td>Study authors’ conflicts of interest</td>
</tr>
<tr>
<td>Ethical approval</td>
</tr>
</tbody>
</table>

Percentages may not add-up to 100% because of rounding errors.
*Also includes reports that explicitly mention that no duplicates were identified.
†Also includes reports that explicitly mention that none of the included studies lacked ethical approval.
regarding the endpoint ‘search for unpublished trials or test for publication bias’. One reviewer declared that, contrary to our assumption, unpublished trials had not been searched in their review, and seven claimed that unpublished trials had been searched although this was not reported in the published reviews.

Eleven reviewers (14%) provided information that changed our data extraction regarding the endpoint ‘contact with authors of original studies’. One declared that authors had not been contacted, and 10 claimed that authors had been contacted although this was not reported in the published reviews.

Twenty-six reviewers (32%) provided information that changed our data extraction regarding the endpoint ‘duplicate publication’. Terms used included ‘duplicate’, ‘companion article’, ‘multiple publications’, ‘articles with overlapping datasets’, or ‘trials with identical patient population’. Three reviewers declared that, contrary to our assumption, they had not identified duplicate publications, and 23 claimed having searched for duplicates although this was not reported in the published review.

Five reviewers (6%) told us about suspected cases of misconduct that were not reported in the published review.

Characteristics of the reviews did not differ depending on whether or not we were able to contact their authors (table 2).

**Main results**

The median number of procedures applied in each review was 2.5 (IQR, 1–3). Eleven reviews (9%) applied no procedures at all, while no review applied all six procedures.

**Search of unpublished trials and test for publication bias**

Fifty-six reviewers (47%) either searched for unpublished trials or applied a statistical test to identify publication bias. Twenty-three reviewers (19%) did both. Unpublished studies were sought for in trial registries (eg, ClinicalTrials.gov or FDA database), or by contacting experts and manufacturers. The number of unpublished studies included in these reviews was inconsistently reported. Contacting the reviewers did not help us clarify this issue. Seven reviews (6%) only discussed the risk of publication bias, and 32 reviews (27%) did not mention it at all (table 3).

**Contact with authors to unearth unreported outcomes**

Seventy-three reviewers (62%) had contacted the authors of the original studies. Fifty-eight reviewers (49%) had searched for unreported results from the original articles. The reviews rarely reported on the number of authors contacted and the response rate. We were not able to clarify this issue in our email exchange with the reviewers (table 3).

**Duplicate publications**

Duplicate publications were sought for in 81 reviews (69%). Twenty-two reviewers confirmed that duplicates were not sought for, and 15 did not answer our enquiry. The number of duplicates identified was rarely mentioned. We failed to clarify this issue in our exchange with the reviewers. Ten reviews (8.5%) published the reference of at least one identified duplicate.

**Sponsors**

Twenty-seven reviews (23%) reported on the sources of funding for the studies. Six reviewers (5%) analysed the
impact of sponsors on the results of the review.\textsuperscript{B(4,15,26,32),J12,L8} One reviewer claimed that sponsor bias was unlikely,\textsuperscript{L8} while three were unable to identify any sponsor bias.\textsuperscript{B(4,15),J12} Finally, two reviews identified sponsor bias (see online supplementary appendix table 3A).\textsuperscript{B26,32}

### Conflicts of interest of authors

Five reviewers reported on conflicts of interest of the authors of the studies.\textsuperscript{A(8,14),B20,C(8,14)} None of them used this information to perform subgroup analyses. One review mentioned conflicts of interest as a possible explanation for their (biased) findings.\textsuperscript{A8} In three reviews, the affiliations of the authors were summarised and potential conflicts of interest clearly identified.\textsuperscript{B20,C(8,14)} Finally, the appendix table of one of the reviews showed that one study might have suffered from a ‘significant conflict of interest’.\textsuperscript{A4}

### Ethical approval of the studies

Three reviews looked at whether or not ethical approval had been sought for (see online supplementary appendix table 3A).\textsuperscript{B(27,37),C10} Two reviews explicitly reported that all included studies had received ethical approval.\textsuperscript{B(27,37)} The third review reported extensively on which studies had or had not provided any information on ethical approval or patient consent.\textsuperscript{C10}

Outcomes did not differ according to whether or not we were able to contact the reviewers (table 3).

### Table 2 Characteristics of the systematic reviews analysed

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>Answer</th>
<th>No answer</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Per cent</td>
<td>N</td>
<td>Per cent</td>
</tr>
<tr>
<td>Number of systematic reviews</td>
<td>118</td>
<td>100</td>
<td>80</td>
<td>68</td>
</tr>
<tr>
<td>Number of co-authors</td>
<td></td>
<td></td>
<td>6 (5–9)</td>
<td>6 (4–7.5)</td>
</tr>
<tr>
<td>Country of affiliation of 1st author</td>
<td>USA</td>
<td>42</td>
<td>36</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>UK</td>
<td>23</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>21</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Australia</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>28</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>Protocol registration/accessibility</td>
<td>No mention of protocol</td>
<td>82</td>
<td>69</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Protocol exists/not freely accessible</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Protocol exists/freely accessible</td>
<td>33</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>Source of funding of systematic review</td>
<td>Not reported</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Reported that no funding was received</td>
<td>24</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Reported that funding was received with details</td>
<td>86</td>
<td>73</td>
<td>58</td>
</tr>
<tr>
<td>Conflicts of interests of systematic reviewers</td>
<td>Not reported</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Reported that there were none</td>
<td>42</td>
<td>36</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Reported and detailed in the published report</td>
<td>61</td>
<td>52</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Reported elsewhere (eg, online)</td>
<td>15</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Number of databases searched</td>
<td>Median (IQR)</td>
<td>4 (3–7)</td>
<td>4 (3–7)</td>
<td>4 (3–6)</td>
</tr>
<tr>
<td>Language limitations applied</td>
<td>None</td>
<td>56</td>
<td>47</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>English only</td>
<td>42</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Limitations to other languages</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Not specified</td>
<td>13</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>Number of studies included</td>
<td>Median (IQR)</td>
<td>28 (12–57)</td>
<td>28 (12–59)</td>
<td>34 (12–55)</td>
</tr>
<tr>
<td>Study designs examined</td>
<td>RCT only</td>
<td>46</td>
<td>39</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Cohort prospective only</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Diagnostic studies only</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Various designs</td>
<td>62</td>
<td>53</td>
<td>39</td>
</tr>
</tbody>
</table>

Answer: reviews in which extracted data were confirmed by reviewers. No answer: reviews in which extracted data were not confirmed by reviewers. p Value testing the null hypothesis of equal distribution between the reviews for which the author responded to our inquiries and those who did not. Statistical tests: \(\chi^2\) test or Kruskal-Wallis equality of population rank tests, as appropriate.

RCT, randomised controlled trials.

\cite{EliaN}
Suspicion of misconduct

Two reviewers suspected research misconduct in the articles included in their review and reported it accordingly.\cite{31,12} Contacting the other reviewers allowed us to uncover five additional cases of possible misconduct. Four agreed to be cited here;\cite{26,33,16,1} while one preferred to remain anonymous.

Data falsification was suspected in three reviews.\cite{26,36,1} One review, looking at the association of hydroxyethyl starch administration with mortality or acute kidney injury of critically ill patients,\cite{12} included seven articles co-authored by Joachim Boldt. However, a survey performed in 2010, and focusing on Boldt’s research published between 1999 and 2010, had led to the retraction of 80 of his articles due to data fabrication and lack of ethical approval.\cite{20} The seven articles co-authored by Boldt were kept in the review as they had been published before 1999. Nonetheless, the reviewers performed sensitivity analyses excluding these seven articles, and showed a significant increase in the risk of mortality and acute kidney injury with hydroxyethyl starch solutions that was not apparent in Boldt’s articles.

The second review examined different techniques of sperm selection for assisted reproduction. The reviewers suspected data manipulation in one study since its authors reported non-significant differences between the number of oocytes retrieved and embryos transferred while the p-value, when recalculated by the reviewers, was statistically significant.\cite{16}

The third review (on management strategies for asymptomatic carotid stenosis) reported that misconduct had been suspected based on the ‘differences in data between the published SAPPGIRE trial and the re-analysed data posted on the FDA website’.\cite{26} Although this information was not provided in the published review, it was available in the full report online (http://www.ahrq.gov/research/findings/ta/carotidstenosis/carotidstenosis.pdf).

| Table 3 | Application of the procedures to counter-balance some common research malpractices |
|--------|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|        | ALL | Answer | No answer | p Value |
| Number of systematic reviews | 118 | 100 | 80 | 68 | 38 | 32 | 0.914 |
| Search of unpublished trials and/or test for publication bias | | | | | | | |
| Publication bias discussed only or not mentioned | 39 | 33 | 26 | 33 | 13 | 34 | |
| Unpublished trials searched OR publication bias tested | 56 | 47 | 39 | 49 | 17 | 45 | |
| Unpublished trials searched AND publication bias tested | 23 | 19 | 15 | 19 | 8 | 21 | |
| Contact with authors of the studies | | | | | | | |
| Study authors not contacted | 45 | 38 | 28 | 35 | 17 | 45 | 0.427 |
| Study authors contacted for method or unspecified reason | 15 | 13 | 12 | 15 | 3 | 8 | |
| Study authors contacted for unreported outcomes | 58 | 49 | 40 | 50 | 18 | 47 | |
| Duplicate publications | | | | | | | |
| Not searched or not mentioned | 37 | 31 | 21 | 26 | 16 | 42 | 0.057 |
| Searched and found, not referenced OR no mention of results | 71 | 60 | 54 | 68 | 17 | 45 | |
| Searched, found and referenced | 10 | 8 | 5 | 6 | 5 | 13 | |
| Sponsors of the studies | | | | | | | |
| Not mentioned | 91 | 77 | 63 | 79 | 28 | 74 | 0.809 |
| Information extracted | 21 | 18 | 13 | 16 | 8 | 21 | |
| Information extracted and subgroup analyses performed | 6 | 5 | 4 | 5 | 2 | 5 | |
| Conflicts of interests of study authors | | | | | | | |
| Not mentioned | 113 | 96 | 77 | 96 | 36 | 95 | 0.703 |
| Information extracted | 5 | 4 | 3 | 4 | 2 | 5 | |
| Information extracted and subgroup analyses performed | 0 | 0 | 0 | 0 | 0 | 0 | |
| Ethical approval of included studies | | | | | | | |
| Not mentioned | 115 | 97 | 77 | 96 | 38 | 100 | 0.481 |
| Information extracted | 3 | 3 | 3 | 4 | 0 | 0 | |
| Information extracted and subgroup analyses performed | 0 | 0 | 0 | 0 | 0 | 0 | |
| Number of procedures applied | | | | | | | |
| None | 11 | 9 | 6 | 8 | 5 | 13 | 0.403 |
| 1 or 2 procedures | 48 | 41 | 34 | 43 | 14 | 37 | |
| 3 or 4 procedures | 56 | 47 | 38 | 48 | 18 | 47 | |
| 5 procedures | 3 | 3 | 2 | 3 | 1 | 3 | |
| Median (IQR) | 2.5 (1–3) | 2.5 (2–3) | 2.5 (1–3) | | | | |
| Explicit mention of misconduct by reviewers | | | | | | | |
| No, or not mentioned | 111 | 94 | 74 | 93 | 37 | 97 | 0.296 |
| Yes | 7 | 6 | 6 | 8 | 1 | 3 | |

Answer: reviews in which extracted data were confirmed by reviewers. No answer: reviews in which extracted data were not confirmed by reviewers. p Value testing the null hypothesis of equal distribution between the reviews for which the authors responded to our inquiry and those who did not (χ² test). Percentages may not add-up to 100% because of rounding errors.
Intentional selective reporting of outcomes was suspected in two reviews. In one review that examined early interventions to prevent psychosis, the reviewers identified, discussed and referenced three articles that did not report on all the outcomes. In the second review, the corresponding reviewer (who preferred to remain anonymous) revealed that he ‘knew of two situations in which authors knew what the results showed if they used standard categories (of outcome) but did not publish them because there was no relationship, or not the one they had hoped to find.’

Plagiarism was identified in one review examining the epidemiology of Alzheimer’s disease and other forms of dementia in China. According to the reviewers, they had identified a ‘copy-paste-like duplicate publication of the same paper published two or more times, with the same results and sometimes even different authors’. This information was not reported in the published review. The corresponding reviewer explained that they had ‘a brief discussion...about what to do about those findings and whether to mention them in the paper...We did not think that we should be distracted from our main goal, so we felt that it was better to leave it to qualified bodies and specialised committees on research malpractice to address this problem separately.’

Finally, one reviewer told us that ‘there were some suspected misconduct’ in original studies on which we performed sensitivity analysis (best case—worst case)’. The review mentioned that some studies were of poor quality, but it did not specifically mention suspicion of misconduct.

The median number of studies included in the reviews that detected misconduct was 56 (IQR, 11–97), and was 28 (IQR, 12–57) in the reviews that did not detect misconduct. The difference did not reach statistical significance.

**Secondary endpoints**

The reviews published in the four medical journals differed in most characteristics examined in this paper. The reviews published in the Cochrane Library differed from all the other reviews (see online supplementary appendix table 2A). There were also differences in procedures applied across the journals (see online supplementary appendix table 4A). The only three reviews that had extracted data on ethical approval for the studies were published in the BMJ (2) and in the Cochrane Library (1). Finally, one review from the BMJ and one review from JAMA explicitly mentioned potential misconduct.

**DISCUSSION**

**Statement of principal findings**

This analysis confirms some issues and highlights new ones. The risk related to double counting of participants due to duplicate publications and the risk of selective reporting of outcomes are reasonably well recognised. More than half of the reviews applied procedures to reduce the impact of these malpractices. The problem of conflicts of interest remains underestimated, and ethical approval of the original studies is overlooked. Although systematic reviewers are in a privileged position to unearth misconduct such as copy-paste-like plagiarism, intentional selective data reporting and data fabrication or falsification, they do not systematically report them. Finally, editors have a role to play in improving and implementing rigorous procedures for the reporting of systematic reviews to counter-balance the impact of research malpractice.

**Comparison with other similar analyses**

Our study confirms that systematic reviewers are able to identify publications dealing with the same cohort of patients. However, 20% of reviews under consideration failed to report having searched for duplicates. Only ten of them provided the references to some of the identified duplicates. It remains unclear whether reviewers do not consider duplicate publication worth disclosing or whether they are unsure on how to address the issue. Finally, there is no widely accepted definition of the term ‘duplicate’, which, in turn, adds to the confusion. For example, a number of reviewers used the term ‘duplicate’ to describe identical references identified more than once through the search process.

Selective publication of studies, and selective reporting of outcomes, have been examined previously. This led the BMJ to call for ‘publishing yet unpublished completed trials and/or correcting or republishing misreported trials’. Other ways to address this issue include registration of the study protocols, searching for unpublished trials, and contacting authors to retrieve unreported outcomes. Our analyses show that 70% of systematic reviewers are aware of these malpractices, although 10% failed to report them explicitly. As described before, most reviewers did not report on the number of unpublished articles included in their analyses, the number of authors contacted and the response rate.

Despite the obvious risk of research conclusions favouring a sponsor, subgroup analyses on funding were rarely performed. Sponsor bias may overlap with other malpractices such as selective reporting, redundant publication or failure to publish completed research. It may also overlap with conflicts of interest of the authors of the original studies, an issue that remains largely overlooked in these systematic reviews. There is a general understanding that authors with conflicts of interest are likely to present conclusions in favour of their own interests. Although most journals now ask for a complete and overt declaration of conflicts of interest from all authors, this crucial information remains unclearly reported. This may explain why we found no reviews that performed subgroup analyses on this issue.

Ten years ago, Weingarten proposed that ethical approval of studies should be checked during the
process of systematic reviewing. However, our study shows that only three reviews reported having done so. The need to report ethical approval in original studies has only recently been highlighted. A case of massive fraud in anaesthesiology, in which informed consent from patients and formal approval by ethics committee were fabricated, only shows how difficult a task it will be.

The most striking finding was that although seven systematic reviews suspected misconduct in original studies, five of them did not report it, one reported it without further comment and only one reported overtly on the suspicion. This illustrates that reviewers do not consider themselves entitled to make an allegation, although they are in a privileged position to identify misconduct. The fact that one reviewer preferred to remain anonymous further illustrates the reluctance to openly report on misconduct.

Strengths and weaknesses

We used a clear and reproducible sampling method that was not limited to any medical specialty. The data analysed had been confirmed by the reviewers. This allowed us to quantify the proportion of procedures that were implemented but not reported by the reviewers. Finally, to our knowledge, this is the first analysis of the procedures used by systematic reviewers to deal with research malpractice and misconduct. Qualitative answers of the reviewers were very informative.

We selected systematic reviews from four major medical journals and the Cochrane Library, which is considered the gold standard in terms of systematic reviewing. We can reasonably assume that the problems identified are at least as serious in other medical journals. Systematic reviews that were not identified as such in their titles were not included. However, including these reviews would not have changed our findings. Only one reminder was sent to the reviewers, since the response rate was reasonably high. Furthermore, the characteristics of the systematic reviews did not differ between reviews for which authors responded or not. We did not ask the reviewers to confirm all their data but focused on the information that we considered unclear. It is possible that some of the reviewers had indeed extracted information on sponsors’ and authors’ conflicts of interest, as well as ethical approval of the studies, but failed to report them. A number of procedures applied and instances of misconduct came to light through our personal contacts with the reviewers. Our results might therefore be underestimated. On the other hand, it is possible that some reviewers might have pretended to have applied some procedures although they had not. This would have led to an overestimation of the number of procedures applied. The major weakness of this study lies in the lack of an accepted definition of ‘research misconduct’. It is possible that some reviewers might have hesitated to disclose suspected misconduct, leading to an underestimation of the prevalence of misconduct identified. Finally, we have used the ‘preliminary taxonomy of research misconduct’ proposed by Smith in 2000, and categorised all common minor misconduct as ‘malpractices’. Some may disagree with our classification.

Conclusions and research agenda

The PRISMA guideline has improved the reporting of systematic reviews. PRISMA-P aims to improve the robustness of the protocols of systematic reviews. The 17-items list mentions the assessment of meta-bias(es), such as publication bias across studies, and selective outcome reporting within studies. However, the list is not concerned with authors’ conflicts of interest, sponsors, ethical approval of original studies, duplicate publications or the reporting of suspected misconduct. The MECIR project defines 118 criteria classified as ‘mandatory’ or ‘highly desirable’, to ensure transparent reporting in a Cochrane review. Among these criteria, publication and outcome reporting bias, funding sources as well as conflicts of interest are highlighted as ‘mandatory’ to report on. However, ethical approval for studies is not. Most importantly, neither of the two recommendations explicitly describes what should be considered misconduct, what kind of misconduct must be reported, to whom, and how.

We have previously shown how systematic reviewers can test the impact of fraudulent data on systematic reviews, identify redundant research and identify references that should have been retracted. This paper suggests that systematic reviewers may have additional roles to play. They may want to apply specific procedures to protect their analyses from common malpractices in the original research, and they may want to identify and report on suspected misconduct. However, they do not seem to be ready to act as whistle-blowers. The need for explicit guidelines on what reviewers should do once misconduct has been suspected or identified has already been highlighted. These guidelines remain to be defined and implemented. The proper procedure would require the reviewer to request the institution where the research was conducted to investigate on the suspected misconduct, as the institution holds the legal legitimacy. Whether alternative procedures could be applied should be discussed. For example, they may include contacting the editor-in-chief of the journal where the suspected paper was originally published, or the editor-in-chief where the systematic review will eventually be published. Future research should explore the application of additional protective procedures such as checking for the adherence of each study to its protocol, or the handling of outlier results, and quantify the impact of these measures on the conclusions of the reviews. Finally, potential risks of false reporting of misconduct need to be studied.

Author affiliations

1Division of Anaesthesiology, Geneva University Hospitals, Geneva, Switzerland

Acknowledgements  The authors thank all the authors of the systematic reviews who kindly answered our inquiries, and Liz Wagner for her thoughtful advice on the first draft of our manuscript.

Contributors  NE, MRT and EV designed and coordinated the investigations. NE wrote the first draft of the manuscript. AC and DMP were responsible for checking the inclusion and extraction of the data. NE performed the statistical analyses and contacted the reviewers. All the authors contributed to the revision of the manuscript and the intellectual development of the paper. NE is the study guarantor.

Funding  This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests  All the authors have completed the Unified Competing Interests form at http://www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that they have received no support from any organisation for the submitted work, that they have no financial or non-financial interests that may be relevant to the submitted work, nor do their spouses, partners, or children. EV is the co-director of Cochrane Switzerland—a Cochrane entity involved in producing, disseminating and promoting Cochrane reviews.

Provenance and peer review  Not commissioned; externally peer reviewed.

Data sharing statement  No additional data are available.

Open Access  This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES

1. Deer B. How the case against the MMR vaccine was fixed. BMJ 2011;342:c5347.


5.5 Ability of a meta-analysis to prevent redundant research: systematic review of studies on pain from propofol injection

Authors: Habre C, Tramer MR, Popping DM, Elia N
Published in: BMJ 2014; 349 :g5219
IF 2014: 17.445

Summary

In this study we aimed to check the potential impact of a systematic review on the performance and design of subsequent research, based on a very well known example from all anaesthesiologists: the problem of prevention of pain on injection of propofol. Specifically, we aimed to verify that the conclusion and recommendations of a systematic review published in 2000 on this topic had been followed. The systematic review provided a clear research agenda stipulating that further trials needed to focus on children and that, since an effective intervention had been identified to prevent pain on injection of propofol, this “gold standard intervention” needed to be used as a comparator to test the clinical relevance of new interventions.

We showed that the impact of the systematic review on the designs of subsequent research was low. In particular, only 49 of 136 trials published after the publication of the systematic review used the gold standard intervention as comparator or were performed in children and could therefore be considered clinically relevant since potentially leading to a change in practice.

This study highlighted the problem of the profusion of clinically irrelevant studies due to the imbalance between the strong pressure to publish and the weak barriers to prevent useless research and calls for a strengthening of existing control barriers such as ethical committees, funders, clinical trial registries and journal editors.

Finally, the article concludes with this new idea regarding the role that systematic reviews ought to play in helping researchers to improve their future research.
Ability of a meta-analysis to prevent redundant research: systematic review of studies on pain from propofol injection

Céline Habre research fellow1, Martin R Tramèr professor in anaesthesia2,3, Daniel M Pöpping anaesthetist4, Nadia Elia public health epidemiologist2,5

1Department of Radiology, Geneva University Hospitals, 4 rue Gabrielle-Perret-Gentil, CH-1211 Geneva 14, Switzerland; 2Division of Anaesthesiology, Geneva University Hospitals, Geneva, Switzerland; 3Faculty of Medicine, University of Geneva, Geneva, Switzerland; 4Department of Anaesthesiology and Intensive Care, University Hospital Münster, Münster, Germany; 5Institute of Global Health, Faculty of Medicine, University of Geneva, Geneva, Switzerland

Abstract

Objective To examine whether, according to the conclusions of a 2000 systematic review with meta-analysis on interventions to prevent pain from propofol injection that provided a research agenda to guide further research on the topic, subsequently published trials were more often optimally blinded, reported on children, and used the most efficacious intervention as comparator; and to check whether the number of new trials published each year had decreased and whether the designs of trials that cited the review differed from those that did not.

Study design Systematic review comparing old trials (published before, and included in, the review) with new trials (published afterwards).

Data sources Medline, Cochrane, Embase, and bibliographies to January 2013.

Eligibility criteria for study selection Randomised studies testing any intervention to prevent pain from propofol injection in humans.

Results 136 new trials (19 778 patients) were retrieved. Compared with the 56 old trials (62 664 patients), the proportion of optimally blinded trials had increased from 10.7% to 38.2% (difference 27.5%, 95% confidence interval 16.0% to 39.0%, P<0.001), and the proportion of trials that used the most efficacious intervention as comparator had increased from 12.5% to 27.9% (difference 15.4%, 4.0% to 26.9%, P=0.022). The proportion of paediatric trials had increased from 5.4% to 12.5%, although this was not significant (difference 7.1%, −1.0% to 15.2%, P=0.141). The number of new trials published each year was significantly higher (median number/year 12 (range 7-20) v 2.5 (0-9), P<0.001) with no obvious decreasing trend. 72.8% (n=99) of the new trials were considered clinically relevant since they used the most efficacious intervention as comparator or included a paediatric population.

Conclusions The impact of the systematic review on the design of subsequent research was low. There was an improvement in the reporting of optimal blinding procedures and a tendency towards an increase in the proportion of paediatric trials. The most efficacious intervention was more often chosen as comparator but remained marginally used, and the number of trials published per year had not decreased. The use of systematic reviews should be encouraged to inform rational, and thus ethical, trial design and improve the relevance of new research.

Introduction

Systematic reviews often identify gaps in knowledge and methodological flaws in the existing literature. They should also guide researchers in assessing the need for further investigations, although this is often overlooked. For example, a systematic review published in 2005 and including 64 trials found that in 1992 it could already have been shown, after the 12th trial had been published, that aprotinin reduced the risk of bleeding in patients undergoing cardiac surgery; thus a timely performed systematic analysis of the published literature could have prevented 52 further trials from being performed.

It remains unclear though to what extent published systematic reviews influence the design of subsequent trials. Designing a trial comprises, among other things, the choice of population and outcomes of interest, methods of data collection,
or a comparator against which a new, potentially useful experimental intervention ought to be tested. Administering the anaesthetic propofol intravenously may be distressing for patients as it is often associated with pain at the injection site. In 2000, a systematic review by Picard and Tramér (a coauthor of the present analysis), including data from 6264 patients from 56 randomised placebo controlled trials, tested the analgesic efficacy of interventions to prevent the pain from propofol injection. The systematic review, which was published in one of the top five anaesthesiology journals, indexed in all major medical databases, provided six main messages. Firstly, evidence showed that the most efficacious analgesic intervention was to administer a small intravenous dose of lidocaine (lignocaine) with venous occlusion before the propofol injection; with lignocaine 40 mg, the best documented regimen, the number needed to treat to prevent any pain compared with placebo was 1.8 (95% confidence interval 1.5 to 2.2). Secondly, although alternative interventions were also efficacious (for example, an intravenous bolus of lignocaine without venous occlusion, lignocaine mixed with propofol, or a variety of opioids or non-opioid analgesics administered concomitantly), none of these showed a similar degree of efficacy compared with lignocaine with venous occlusion (see supplementary file 1). Thirdly, for several experimental interventions, such as intravenous ondansetron, droperidol, or ketamine, or the dilution of propofol with homologous blood, no meaningful conclusions could be drawn owing to a lack of valid data. Fourthly, more data on children were needed to allow definite conclusions in this population. Fifthly, blinding procedures needed to be improved as only 11% of the trials were optimally blinded. And finally the authors of the Picard review concluded their report by questioning the necessity of performing further trials to identify yet another analgesic intervention to prevent pain from propofol injection.

We examined whether the Picard review had had any impact on subsequent research on pain from propofol injection. Specifically, we checked whether the number of new trials on the subject published per year had decreased over time, and whether the Picard review had influenced the design of subsequently published trials. For example, we expected that most new trials would focus on children, that the proportion of optimally blinded trials would increase, and that the most efficacious analgesic intervention identified in the Picard review would be chosen as a comparator against which new experimental interventions were tested. We also checked whether subsequently published trials cited the Picard review, and whether the designs of trials that cited the review differed from those that did not.

Methods
This systematic review was written according to the PRISMA statement for reporting systematic reviews and meta-analyses. The study protocol was not registered but is available from the authors.

Eligibility criteria
We included all trials that had been analysed in the Picard review and added a new search to identify all trials that had since been published. As in the Picard review, we searched for full published reports of randomised trials testing the analgesic efficacy of any intervention compared with placebo or no treatment to prevent pain from propofol administered intravenously. Since the necessity (and ethical acceptability) of a placebo arm may be questioned as an efficacious intervention had been identified, we additionally searched for trials that did not include a placebo arm. We included trials in adults, children, or volunteers undergoing general anaesthesia or sedation. We considered drug interventions (for example, pretreatment with a drug, or an alternative emulsion of propofol) and non-drug interventions (for example, cooling of propofol). To be included, trials had to report on the incidence of pain on injection of propofol as the primary outcome. We did not consider letters, conference abstracts, or studies in animals.

Information sources and searches
We performed searches in Medline (via Pubmed), Embase, and the Cochrane Library. We additionally identified trials from bibliographies of retrieved trials and checked references of a further relevant systematic review by Jalota and colleagues that was published 11 years after the Picard review. We limited the search period from January 2002 (to ensure that trialists had the scope to read the Picard review, published in 2000) to January 2013. We did not search for unpublished trials. Trials were identified using the same search strategy and key words as in the Picard review—namely, “propofol”, “pain”, “injection”, and “random”, sought in the titles and abstracts, with a limit to humans but no limit to language.

Study selection and risk of bias assessment
One author (CH) assessed the eligibility of retrieved articles by screening the titles and abstracts. Queries were resolved through discussion with two other authors (NE, MRT). As in the Picard review, we scored new trials for quality of data reporting using the five point Oxford scale, which considers the three items randomisation, blinding, and flow of patients. Blinding was rated optimal (2 points) when drugs were matched. Since we analysed exclusively randomised trials, the minimum score of an included trial was 1.

Data collection process
One author (CH) entered all data into an excel spreadsheet, which was developed for the purpose of this analysis. One of three other authors (NE, DMP, or MRT) independently checked the data. We contacted the authors of the original reports when we needed to clarify the nature of the data or were unable to access a report.

Data items
We extracted the characteristics of the trials, including year of publication, journal impact factor (Journals Citation Reports 2011; we analysed journals without an impact factor separately), open access status of the journal (yes/no), study population (adults, children, volunteers), number of analysed participants, and sources of funding (none, academic, industry, not declared). According to the Picard review, the most efficacious intervention was an intravenous bolus injection of lignocaine with venous occlusion (manually or with a tourniquet) about 20 seconds before administering propofol into the same vein. For the purpose of our analysis we classified this method as the primary reference treatment. Alternative interventions that had some proved efficacy, although less so compared with the lignocaine occlusion technique (for instance, intravenous injection of lignocaine without occlusion), were classified by us as secondary reference treatments. We classified interventions without proved efficacy according to the Picard review, and new, potentially useful interventions that had not yet been retrieved in that systematic review, as experimental interventions. We regarded no treatment controls as placebos.
Characteristics of new trials

The 136 new trials included data from 19778 patients (table 1). The trials originated from 30 countries and were published in 51 different journals, of which 29 (56.9%) had an impact factor and 14 (27.5%) were open access (see supplementary file 2). Eighty trials (58.8%) tested the efficacy of a variety of drugs administered before, or concomitantly with, propofol, 40 (29.4%) tested the analgesic efficacy of different emulsions of propofol, and 16 (11.8%) tested non-drug interventions.

Synthesis of differences between old and new trials

General characteristics

Compared with the old trials, the new trials were larger (median number of analysed participants 125.5 v 100, P<0.001), published in journals with lower impact factors (median 2.23 v 2.96, P<0.001), more often published in journals without an impact factor (30.1% v 16.1%, difference 14.1%, 95% confidence interval 1.7% to 26.4%, P=0.043), and scored higher for quality of data reporting (median 3 v 2, P<0.001, table 1). Eighty nine (65.4%) new trials scored 2 for randomisation compared with 6 (10.7%) old trials (difference 54.7%, 95% confidence interval 43.3% to 66.1%, P<0.001). The flow of patients was optimally described in 29 (21.3%) new trials compared with five (8.9%) old trials (difference 12.4%, 2.2% to 22.6%, P=0.041).

Main outcomes

Compared with the old trials, the number of new trials published per year increased (median 12 v 2.5; P<0.001, fig 2, table 2). The number of new trials performed in children had also increased, from three (5.4%) to 17 (12.5%), although the difference in proportions did not reach significance (P=0.141). Blinding procedures had improved (P<0.001) and a greater proportion of new trials were optimally blinded (10.7% v 38.2%; difference 27.5%, 16.0% to 39.0%). New trials used the primary reference treatment more often as a comparator (27.9% v 12.5%; difference 15.4%, 4.0% to 26.9%, P=0.022) and used a secondary reference treatment less often (35.3% v 62.5%; difference −27.2%, −42.2% to −12.2%, P<0.001).

Ninety nine of the 136 new trials (72.8%) cited the Picard review (table 3). Compared with the 37 trials that did not, trials citing the review were published more often per year (median 9 v 3, P<0.001), were not more often performed in children (14.1% v 8.1%, P=0.344), the distribution of blinding scores were not statistically significantly different although optimal blinding was more common (43.4% v 24.3%, difference 19.1%, 2.2% to 36.0%), and reporting quality tended to be higher (median Oxford score 3 v 2, P=0.011). Reporting of randomisation procedures scored 2 in 70 (70.7%) of the trials citing the review compared with 19 (51.4%) of the trials not citing the review (difference 19.4%, 0.9% to 37.8%, P=0.035), and the flow of patients was optimally described in 24 (24.2%) of the trials citing the review compared with five (13.5%) of the trials not citing the review (P=0.174). The proportion of trials including the primary reference treatment did not differ (29.2% v 24.3%, P=0.565, table 4). However, trials citing the review included a secondary reference treatment more often (40.4% v 21.6%; difference 18.8%, 2.4% to 35.2%, P=0.041) and a design without a primary or a secondary reference treatment less often (30.3% v 54.1%; difference −23.8%, −42.2% to −5.3%, P=0.011).

Results

Study selection

We identified 360 new reports (fig 1). Through screening of titles and abstracts, we excluded 189 reports. The remaining 171 were studied in detail and a further 35 were subsequently excluded. We eventually included 136 new randomised trials that had been published at least two years after the publication of the Picard review (see supplementary files 2 and 3). Of these 136 new trials, 94 were placebo controlled.
**Additional findings**

**Clinical relevance**

Forty nine (36.0%) new trials were regarded by us as clinically relevant. Of the 87 (64.0%) trials considered not clinically relevant, 47 (54.0%) compared experimental interventions with or without a placebo, 32 (36.8%) compared an experimental intervention with a secondary reference treatment with or without a placebo, four (4.6%) compared the primary reference treatment with a secondary reference treatment with or without a placebo, three (3.5%) compared secondary reference treatments with or without a placebo, and one (1.1%) compared the primary reference treatment with placebo.

There were no significant differences between clinically relevant and non-relevant trials for the proportion that cited the Picard review (79.6% v 69.0%, P=0.181), the proportion published in journals without an impact factor (28.6% v 31.0%, P=0.764), the median impact factor of those published in journals with an impact factor (2.23 v 2.19, P=0.418), or the proportion published in open access journals (22.4% v 20.7%, P=0.810, table 5).

Of the 49 clinically relevant trials, two (4.1%) were funded by industry; of the 87 clinically non-relevant trials, seven (8.0%) were funded by industry (difference −4.0%, −11.9% to 4.0%, table 5). Of the nine new trials funded by industry, two were optimally blinded, two were not double blinded at all, five (55.5%) compared an experimental intervention with a secondary reference treatment, three (33.3%) compared an experimental intervention with placebo, and one (11.1%) compared two experimental interventions. None used the primary reference treatment (see supplementary file 2).

**Context of citation of Picard review**

Since we were unable to find significant differences between the designs of trials citing and not citing the Picard review, we made further investigations regarding the context in which the Picard review was cited in the original trials. Among the 99 trials that cited the Picard review, 39 (39.4%) did so to illustrate the large variety of analgesic interventions that had been tested in this setting (five used the primary reference treatment) and 21 (21.2%) to document the underlying risk of pain (four used the primary reference treatment). Fifteen trials explicitly reported having used the review as a basis for the choice of the primary reference treatment as a comparator (of which one criticised the fact that the studies included in the review were too disparate and that they had methodological gaps, and therefore they chose to repeat the comparison). Seven trials explicitly reported not having chosen the primary reference treatment for different reasons; the tourniquet was difficult or threatening in children (n=4 trials), the primary reference treatment was not often used although it was the best option available (n=1), the primary reference treatment was judged to be awkward (n=1), and the Picard review was regarded as not-conclusive (n=1). Eleven further trials acknowledged the identification of the primary reference treatment by the review; 10 ignored it without any explanation and one chose the primary reference treatment, but on the basis of the authors’ own previous pilot study. The authors of four trials chose the primary reference treatment without explanation but eventually compared their results with those of the Picard review. Finally, the authors of two trials stated as a limitation of their study that they did not use the primary reference treatment that was identified by the Picard review.

All analyses were performed on the subgroup of 94 trials including a placebo or no treatment group. The results were similar.

**Discussion**

This study, which aimed to examine the impact of a systematic review with meta-analysis on the design and relevance of subsequent research, illustrates four major problems. Firstly, although the systematic review had identified a simple, effective, and low cost intervention and strongly suggested that additional trials on this specific issue were no longer necessary, the publication of trials has not decreased. Secondly, although the systematic review provided a clear research agenda, its influence on the design of further trials has remained poor. Thirdly, the proportion of subsequently published trials that could have had an impact on clinical practice has remained low. Finally, citing the systematic review had no clear influence on the design or relevance of subsequently published research.

**Comparison with other studies**

There are numerous examples where systematic reviews, if performed in a timely manner, could have provided evidence of the effectiveness of an intervention and thus prevented redundant research. There is also evidence that knowledge from systematic reviews is underused to inform future research. Our study confirms these findings. These raise ethical concerns not only because patients are unnecessarily randomised in worthless trials but also because resources are wasted.

It remains unclear why the number of published trials on the prevention of pain from propofol injection has not decreased; in fact the number has actually increased. Individual motivations may explain this finding. Doctors are challenged to engage in research for their career progression (publish or perish policy), but clinically relevant large studies with long follow-up periods are difficult to achieve. Studies focusing on pain from propofol injection are easy and quick to perform and are straightforward to publish. As long as academic promotion and funding systems are based on simplistic counts of published papers, favouring prolific authors regardless of the relevance and validity of their work, this is unlikely to change. Ethics committees are supposed to ensure scientific soundness and relevance of clinical investigations, preventing enrolment of participants in non-ethical trials. This should include the rational choice of a comparator intervention and be supported by systematic reviews to check how new protocols fit in with the current state of medical knowledge. In real life, however, ethics committees seem to stand alone, with limited resources and sometimes not enough scientific credit or knowledge to identify, and stop, the performance of irrelevant research. It also remains unclear why editors accepted these new trials for publication. The new trials were published in journals with lower impact factors, and a higher proportion was published in journals without any impact factor. This suggests that editors of higher quality journals were unwilling to publish articles on an already solved problem. It has been suggested that open access journals may apply less stringent criteria for publication. We cannot confirm this hypothesis based on our sample of trials; a similar proportion of relevant and non-relevant trials were published in open access journals.

The Picard review provided a clear research agenda; however, its influence on the design of further research has remained marginal. More data on children were deemed necessary, and although 17 new paediatric trials have been published, the proportional increase in paediatric trials did not reach statistical significance and may have occurred by chance. Interestingly, although a total of 20 trials performed in children is now available and may be sufficient to draw conclusions on this
population, these trials were excluded from the updated systematic review by Jalota and colleagues, and to our knowledge no systematic review has yet summarised the available evidence on the prevention of pain from propofol injection in children. The research agenda in the Picard review also suggested that more optimally blinded trials were needed, which was achieved in subsequent trials. Moreover, blinding was more often optimal in trials citing than not citing the Picard review. However, since new studies scored higher in all aspects of quality of data reporting, the specific impact of the Picard review may be questioned, since this improvement may actually reflect an increase in the implementation of the CONSORT statement.15 Interestingly, the source of funding was still not declared in about 65% of the trials, and only three have been registered, although these items are included in the CONSORT checklist.

The identification of the currently most effective intervention to prevent pain from injection of propofol was the main message of the Picard review. It is plausible that in some of the subsequently published trials the review had an influence on the choice of the comparator against which a new, potentially innovative, experimental intervention was tested. It has been argued that the choice of a comparator intervention should be supported by a systematic review of the relevant literature.16 However, the authors of less than one third of subsequently published trials chose the primary reference treatment as a comparator.

Overall, the number of clinically relevant trials (those that are likely to have an impact on clinical practice) remained low. Although the Picard review identified a primary reference treatment, researchers may have wanted to test yet another experimental intervention that may have been even more efficacious, or simpler to use. One would expect this experimental intervention to be compared with the current primary reference treatment. In our study, about 1 in 3 subsequently published trials only used the primary reference treatment as a comparator. Authors that aim for academic recognition may embark on research that does not necessarily improve existing knowledge.17 Also, they may want to reach statistically significant results to ensure publication, since journals are more prone to publish such results,18 and therefore refrain from comparing an experimental intervention with the currently most efficacious treatment. Interestingly, nine trials were industry sponsored, and none of those used the primary reference treatment as a comparator. This suggests that in these trials a comparator was chosen to favour a priori the experimental intervention, at the cost of threatening the principle of equipoise.19 Citing the Picard review did not clearly reflect its influence on study design. Although almost 73% of the new trials cited the Picard review, their characteristics and clinical relevance were largely similar to those that did not. Only about 32% of the trials mentioning the review explicitly described having used it to inform study design; 15 cited it to justify the choice of the primary reference treatment, whereas seven cited it to explain why they had chosen not to use the primary reference treatment. Ignoring the recommendations of the Picard review may be justified as long as it is explicitly explained.

Strengths and weaknesses of this study

This study has several limitations. Firstly, we did not analyse some trials that were published in Chinese, Japanese, or Persian, because of problems with translation, and we did not include studies published between January 2000 and December 2001. It is unlikely that including those trials would have changed the results of our analysis. Secondly, we cannot exclude that some authors remained unaware of the Picard review even though we selected trials that had been published from 2002. A two year delay between the publication of the review and the inclusion of trials into our analyses may be short. Seven of 13 trials (54%) published in 2002 cited the review and it is possible that some of these trials had started enrolment of patients before the publication of the review. This would explain why their design was not influenced by the Picard review. Thirdly, the relevance of the Picard review may be questioned since it may be considered out of date. The median survival time of a systematic review has been estimated to be about six years.20 However, the biological basis underlying pain from propofol injection has not changed, and the Picard review has remained the only systematic review on this topic for almost 12 years. The relevance of the primary reference treatment identified in the review may also be questioned since it was based on about 400 patients only. However, to find an even smaller treatment effect (for example, a decrease from a conservative baseline risk of 60% pain with placebo to 20% pain with an experimental intervention, number needed to treat 2.5), only about 35 patients would be necessary in each group of a single trial to obtain 90% power for detecting such a level of analgesic efficacy (two sided test, α level 0.05). In 2011, a new systematic review on the same subject, published by different authors, included data from 25 260 patients from 177 trials and came to the same conclusion—namely, that administering lidocaine (lignocaine) along with venous occlusion was the most efficacious intervention for the prevention of pain from propofol injection. This second review also reported that injection of propofol through a large vein provided a similar degree of pain relief, an intervention that had remained ill described at the time of the Picard review. Fourthly, we did not contact the authors of the new trials to ask them whether they knew about the Picard review, why they had chosen to cite it or not, and on what grounds they had designed their study. It may be that some authors had actually used the Picard review to inform their study design but did not overtly refer to it. It is also possible that the new trials, which were designed according to the Picard review, subsequently influenced further trials too. Indeed, some trials used a clinically relevant study design but did not explicitly justify their choice. Fifthly, our definition of what constitutes a clinically relevant study design may be challenged. For example, seven trials had chosen, based on the Picard review, not to use the primary reference treatment for different reasons. Of these, two were classified by us as being not clinically relevant. There may be an argument to classify those as clinically relevant as the authors used the Picard review to defend the choice of their comparator. And finally, the generalisability of our findings remains unknown. This analysis is based on a specific subject of perioperative medicine. It is possible that in research areas where studies of longer follow-up times and more complex infrastructure are required, fewer irrelevant trials are produced.

Unanswered questions and future research

Our study raises questions about the dissemination of results identified through a systematic review. It seems that dissemination of the main results of the Picard review and their implementation into practice and methodological guidelines has not been sufficient. The extraordinary degree of analgesic efficacy of intravenous lignocaine with venous occlusion, its ease of use, and low cost, may have led the authors of the Picard review to believe that the primary reference treatment would be widely accepted both in clinical practice and as a comparator for subsequent research. They were wrong. It was surprising that so many new trials had been published on a topic for which
an efficacious intervention was known. Even more disturbing was that most new trials ended up with clinically non-relevant designs, which meant that these trials were unable, regardless of the results, to generate important new knowledge and thus to considerably change clinical practice. The profusion of clinically non-relevant studies reflects the considerable imbalance between the strong pressure to publish and the weakness of barriers that prevent useless research. Methodological implementation strategies are needed to define an appropriate and ethical research agenda based on the best currently available evidence. Dissemination and implementation of scientific knowledge is a challenge and has been widely discussed.21,22 There is a difference, however, between implementation of clinical knowledge and implementation of methodological knowledge. While clinicians remain free to treat patients as they want, research protocols have to go through a range of barriers that could avoid waste of resources if they were adequately used. Barriers include funders (for funded research), ethics committees, and clinical trial registries. Unfortunately, clinical trial registries do not yet exert any quality control over registered protocols; this is supposed to be done by funding agencies and ethics committees. However, funders are more easily impressed by a long list of publications than by relevant content, and ethics committees are more concerned about patient protection than down-regulation of non-relevant research, although scientific soundness is a prerequisite of ethical soundness (for example, Helsinki Declaration article 21).23 We suggest that both funders and ethics committees start asking authors explicitly whether a systematic review on their research topic exists, and if yes, whether that systematic review proposes a research agenda. If this is the case, authors should explain how their proposed research project fits into that research agenda. This would require a wider acceptance of the strengths of systematic reviews and meta-analyses in defining research agendas but would avoid unnecessary, redundant, or invalid, and thus unethical, research. Finally, journal editors may inform their readers that they will stop considering reports of trials that ignore such knowledge for publication.

Conclusions and recommendations

Much effort has been put into the conviction of academia and policymakers that systematic reviews are vital instruments to improve efficacy and safety of healthcare.24,25 These efforts should be extended to the research area. Authors should justify their trial design in the context of the current state of knowledge.26,27 Our findings highlight the role that systematic reviews should play in guiding trialists in their choice of the most appropriate study design, avoiding ill designed and clinically none relevant trials and thus a waste of resources.

We thank P Picard (Centre d’Étude et de Traitement de la Douleur, Clermont-Ferrand, France) and M R Tramèr (Geneva University Hospitals, Geneva, Switzerland) who shared with us original data of their systematic review; T Perneger (Geneva University Hospitals) for his thoughtful advice; R Abbasiwash (Department of Anaesthesiology, Faculty of Medicine, Urmia University of Medical Sciences, Urmia, Iran) who responded to our enquiry and provided a copy of his paper; M Öztürk (Department of Surgery, Geneva University Hospitals) who translated two Turkish papers into English. We assume full responsibility for the analyses and interpretation of these data.

Presentation: Summary data of this analysis have been presented as an oral presentation at the 7th International Congress on Peer Review and Biomedical Publication, Chicago, September 8-10, 2013 (abstract No PRC-13-0043).

Contributors: CH, MRT, and NE conceived the study, contributed to study design, conducted the analyses, and prepared the manuscript. CH performed systematic literature searches. CH and DP extracted data from original trials. DMP contributed to the preparation of the manuscript. All authors read and approved the manuscript. CH and NE had full access to all of the data in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. NE is guarantor. Funding: This study did not receive specific funding. Competing interests: All authors completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declared no potential conflicts of interest (author NE is editor of the PLoS review; no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: No additional data available.

Transparency: The lead author (NE) affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies have been explained.

References


Accepted: 3 August 2014

Cite this as: BMJ 2014;349:g5219

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non-Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works under the same terms.
What is already known on this topic

It is unethical to embark on new research without first analysing what can be learnt from existing literature
Systematic reviews guide researchers in assessing the need for further investigations, to avoid unnecessary and redundant research
Systematic reviews often provide a research agenda to guide future research

What this study adds

A systematic review that had identified an efficacious analgesic intervention to prevent pain on injection of propofol and provided a clear agenda for future research had only little impact on the design of subsequently published trials on the same subject
Implementation strategies are needed to ensure dissemination of results and methodological issues identified through systematic reviews
New trials should be designed according to the findings and recommendations of the research agenda of previously published systematic reviews
### Tables

**Table 1** Characteristics of trials published before and after the Picard review. Values are numbers (percentages) unless stated otherwise.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Old trials*</th>
<th>New trials†</th>
<th>Difference in proportions (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials</td>
<td>56</td>
<td>136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>6264</td>
<td>19 778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range) year of publication</td>
<td>1995 (1982-99)</td>
<td>2007 (2002-12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median No (range) of participants per trial</td>
<td>100 (28-368)</td>
<td>125.5 (16-500)</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Published in journal without impact factor</td>
<td>9 (16.1)</td>
<td>41 (30.1)</td>
<td>14.1 (1.7 to 26.4)</td>
<td>0.043</td>
</tr>
<tr>
<td>Median (range) impact factor‡</td>
<td>2.96 (1.21-5.36)</td>
<td>2.23 (0.03-5.36)</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Oxford quality score§:</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1 or 2</td>
<td>43 (76.8)</td>
<td>45 (33.1)</td>
<td>−43.7 (−57.3 to −30.1)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12 (21.4)</td>
<td>43 (31.6)</td>
<td>10.2 (−3.1 to 23.5)</td>
<td></td>
</tr>
<tr>
<td>4 or 5</td>
<td>1 (1.8)</td>
<td>48 (35.3)</td>
<td>33.5 (24.8 to 42.3)</td>
<td></td>
</tr>
</tbody>
</table>

*Published before (and therefore included in) the Picard review.
†Published after the Picard review, from 2002 onwards.
‡Calculated excluding trials that were published without an impact factor.
§Randomisation: none (score 0), mentioned (1), described and adequate (that is, computer generated list of random numbers, or sealed envelopes) (2); blinding: none or observer blinding only (score 0), double blinding described but not optimal (1), optimal (double dummy or convincing blinding procedures) (2); drop-outs: not described or incomplete (score 0), clear follow-up of each patient (1).
Table 2  Comparisons of outcomes in trials published before and after the Picard review. Values are numbers (percentages) unless stated otherwise

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Old trials*</th>
<th>New trials†</th>
<th>Difference in proportions (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication rate:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median No (range) of trials published per year</td>
<td>2.5 (0-9)</td>
<td>12 (7-20)</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Population:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trials in children</td>
<td>3 (5.4)</td>
<td>17 (12.5)</td>
<td>7.1 (-1.0 to 15.2)</td>
<td>0.141</td>
</tr>
<tr>
<td>Blinding (Oxford quality score):</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No attempt, or single blinding only (score 0)</td>
<td>19 (33.9)</td>
<td>37 (27.2)</td>
<td>6.7 (21.2 to 7.8)</td>
<td></td>
</tr>
<tr>
<td>Described but not optimal (score 1)</td>
<td>31 (55.4)</td>
<td>47 (34.6)</td>
<td>-20.8 (-36.1 to -5.5)</td>
<td></td>
</tr>
<tr>
<td>Optimal (score 2)</td>
<td>6 (10.7)</td>
<td>52 (38.2)</td>
<td>27.5 (16.0 to 39.0)</td>
<td></td>
</tr>
<tr>
<td>Trial design including primary reference treatment:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v ST v experimental (v placebo)</td>
<td>3 (5.4)</td>
<td>5 (3.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v experimental (v placebo)</td>
<td>0 (0.0)</td>
<td>28 (20.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v ST (v placebo)</td>
<td>4 (7.1)</td>
<td>4 (2.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v placebo</td>
<td>0 (0.0)</td>
<td>1 (0.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7 (12.5)</td>
<td>38 (27.9)</td>
<td>15.4 (4.0 to 26.9)</td>
<td>0.022</td>
</tr>
<tr>
<td>Trial design without PT but including ST:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST v experimental (v placebo)</td>
<td>19 (33.9)</td>
<td>45 (33.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST v ST (v placebo)</td>
<td>16 (28.6)</td>
<td>3 (2.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35 (62.5)</td>
<td>48 (35.3)</td>
<td>-27.2 (-42.2 to -12.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Trial design without PT or ST:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental v experimental (v placebo)</td>
<td>14 (25.0)</td>
<td>50 (36.8)</td>
<td>11.8 (-2.2 to 25.7)</td>
<td>0.116</td>
</tr>
</tbody>
</table>

PT=primary reference treatment (intravenous lidocaine (lignocaine) with venous occlusion); ST=secondary reference treatment (that is, alternative intervention of proved efficacy, for example, lignocaine added to propofol).

*Published before (and therefore included in) the Picard review.
†Published after the Picard review, from 2002 onwards.
Table 3  | Comparison of characteristics of trials according to their reference to the Picard review. Values are numbers (percentages) unless stated otherwise

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Reference to review</th>
<th>No reference to review</th>
<th>Difference in proportions (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials</td>
<td>99 (72.8)</td>
<td>37 (27.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>14 590</td>
<td>5188</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range) year of publication</td>
<td>2007 (2002-12)</td>
<td>2007 (2002-12)</td>
<td></td>
<td>0.992</td>
</tr>
<tr>
<td>Median No (range) of participants per trial</td>
<td>127 (22-500)</td>
<td>120 (16-335)</td>
<td></td>
<td>0.839</td>
</tr>
<tr>
<td>Published in journal without impact factor</td>
<td>26 (26.3)</td>
<td>15 (40.5)</td>
<td>−14.3 (−32.3 to 3.8)</td>
<td>0.106</td>
</tr>
<tr>
<td>Median (range) impact factor*</td>
<td>2.23 (0.03-5.36)</td>
<td>2.11 (0.52-3.29)</td>
<td></td>
<td>0.385</td>
</tr>
<tr>
<td>Oxford quality score†</td>
<td></td>
<td></td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>1 or 2</td>
<td>26 (26.3)</td>
<td>19 (51.4)</td>
<td>−25.1 (−43.4 to −6.8)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>32 (32.3)</td>
<td>11 (29.7)</td>
<td>2.6 (−14.8 to 20.0)</td>
<td></td>
</tr>
<tr>
<td>4 or 5</td>
<td>41 (41.4)</td>
<td>7 (18.9)</td>
<td>22.5 (6.6 to 38.4)</td>
<td></td>
</tr>
</tbody>
</table>

*Calculated excluding trials that were published without an impact factor.
†Randomisation: none (score 0), mentioned (1), described and adequate (that is, computer generated list of random numbers, or sealed envelopes) (2); blinding: none or observer blinding only (score 0), double blinding described but not optimal (1), optimal (double dummy or convincing blinding procedures) (2); drop-outs: not described or incomplete (score 0), clear follow-up of each patient (1).
<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Reference to review</th>
<th>No reference to review</th>
<th>Difference in proportions (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication rate:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median No (range) of trials published per year</td>
<td>9 (6-15)</td>
<td>3 (0-6)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Population:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trials in children</td>
<td>14 (14.1)</td>
<td>3 (8.11)</td>
<td>6.0 (-5.1 to 17.2)</td>
<td>0.344</td>
</tr>
<tr>
<td>Blinding (Oxford quality score):</td>
<td></td>
<td></td>
<td></td>
<td>0.118</td>
</tr>
<tr>
<td>No attempt, or single blinding only (score 0)</td>
<td>24 (24.2)</td>
<td>13 (35.1)</td>
<td>-10.9 (-28.4 to 6.6)</td>
<td></td>
</tr>
<tr>
<td>Described but not optimal (score 1)</td>
<td>32 (32.3)</td>
<td>15 (40.5)</td>
<td>-8.2 (-26.5 to 10.1)</td>
<td></td>
</tr>
<tr>
<td>Optimal (score 2)</td>
<td>43 (43.4)</td>
<td>9 (24.3)</td>
<td>19.1 (2.2 to 36.0)</td>
<td></td>
</tr>
<tr>
<td>Trial designs including PT:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v ST v experimental (v placebo)</td>
<td>3 (3.0)</td>
<td>2 (5.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v experimental (v placebo)</td>
<td>23 (23.2)</td>
<td>5 (13.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v ST (v placebo)</td>
<td>3 (3.0)</td>
<td>1 (2.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT v placebo</td>
<td>0 (0.0)</td>
<td>1 (2.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29 (29.2)</td>
<td>9 (24.2)</td>
<td>5.0 (-11.5 to 21.5)</td>
<td>0.565</td>
</tr>
<tr>
<td>Trial designs without PT but including ST:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST v experimental (v placebo)</td>
<td>37 (37.4)</td>
<td>8 (21.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST v ST (v placebo)</td>
<td>3 (3.0)</td>
<td>0 (0.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40 (40.4)</td>
<td>8 (21.6)</td>
<td>18.8 (2.4 to 35.2)</td>
<td>0.041</td>
</tr>
<tr>
<td>Trial designs without PT or ST:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental v experimental (v placebo)</td>
<td>30 (30.3)</td>
<td>20 (54.1)</td>
<td>-23.8 (-42.2 to -5.3)</td>
<td>0.011</td>
</tr>
</tbody>
</table>

PT=primary reference treatment (intravenous lidocaine (lignocaine) with venous occlusion). ST=secondary reference treatment (that is, alternative intervention of proved efficacy, for example, lignocaine added to propofol).
### Table 5  Relevant versus non-relevant trial designs among 136 new trials. Values are numbers (percentages) unless stated otherwise

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Relevant designs</th>
<th>Non-relevant designs</th>
<th>Difference in proportions (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials</td>
<td>49 (36.0)</td>
<td>87 (64.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range) year</td>
<td>2007 (2002-12)</td>
<td>2007 (2002-12)</td>
<td></td>
<td>0.667</td>
</tr>
<tr>
<td>Reference to Picard review</td>
<td>39 (79.6)</td>
<td>60 (69.0)</td>
<td>10.6 (−4.3 to 25.5)</td>
<td>0.181</td>
</tr>
<tr>
<td>Published in journal without impact factor</td>
<td>14 (28.6)</td>
<td>27 (31.0)</td>
<td>−2.4 (−18.4 to 13.5)</td>
<td>0.764</td>
</tr>
<tr>
<td>Median (range) impact factor*</td>
<td>2.23 (0.03-5.36)</td>
<td>2.19 (0.32-4.24)</td>
<td></td>
<td>0.418</td>
</tr>
<tr>
<td>Published in open access journal</td>
<td>11 (22.4)</td>
<td>18 (20.7)</td>
<td>1.8 (−12.7 to 16.2)</td>
<td>0.810</td>
</tr>
<tr>
<td>Funding source:</td>
<td></td>
<td></td>
<td>0.143</td>
<td></td>
</tr>
<tr>
<td>Not declared</td>
<td>32 (65.3)</td>
<td>55 (63.2)</td>
<td>2.1 (−14.7 to 18.8)</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>10 (20.4)</td>
<td>8 (9.2)</td>
<td>11.2 (−1.6 to 24.0)</td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>5 (10.2)</td>
<td>17 (19.5)</td>
<td>−9.3 (−21.2 to 2.5)</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>2 (4.1)</td>
<td>7 (8.0)</td>
<td>−4.0 (−11.9 to 4.0)</td>
<td></td>
</tr>
</tbody>
</table>

*Calculated excluding trials that were published without an impact factor.
**Figures**

**Fig 1** Flow chart of study selection. *Number does not add up as some titles were from more than one database. †Chinese (n=3 trials), Japanese (n=7), and Persian (n=1)*

**Fig 2** Number of published randomised controlled trials studying efficacy of interventions for prevention of pain from propofol injection. Trials published before 2000 are those included in the Picard review. For the present analysis, searches for trials published after the Picard review included references from 2002 onwards. White bars represent clinically relevant trials.
6 Discussion

Medicine has slowly moved from an art to a science based on evidence, and most of today’s clinical practices need to be justified by scientific evidence.\textsuperscript{146,147} This transition aimed to improve treatment efficacy and patient’s safety, and to allow physicians to inform and advise their patients in a clear and transparent way.\textsuperscript{148} The use, and the publication, of scientific evidence has become part of an academic physician’s everyday work, and almost a mandatory task in academic hospitals.\textsuperscript{149} In anaesthesiology, publication of an article is now also required to achieve a specialist degree.\textsuperscript{150} This shift towards science had different consequences of which I will briefly discuss three.

One consequence has been the emergence of a “pressure to publish” atmosphere, which leads to a high increase in the number of articles published in all disciplines, making it almost impossible for physicians to remain up-to-date.\textsuperscript{151,152} This has highlighted an urgent need for clear and transparent reviews of the existing literature, based on scientific and reproducible methods to synthetize previously performed researches, now known as systematic reviews.\textsuperscript{110,147} A second consequence was that, since sound evidence requires implementation of clinical trials using very specific procedures, some of which remain unknown to physicians performing the trials, a large part of the evidence published is of poor quality and may report biased findings.\textsuperscript{152,153,154} Finally, the explosive association of pressure to publish on one side, and of low methodological knowledge on the other, has created the bases of an increased temptation to “cheat” or “fraud”,\textsuperscript{15,155} in order to facilitate the publication of articles. This has further led to the publication of distorted results and conclusions\textsuperscript{156} with potentially devastating consequences on patient’s health,\textsuperscript{157} the scientific knowledge, and the reputation of science as described above.

All of these problems have been reasonably well recognized throughout the medical community and among editors of scientific journals.\textsuperscript{158,159,160} Although Universities fear the scandal related to the uncovering of research fraud in their institution,\textsuperscript{161} and remain largely reluctant to openly communicate on these issues, once identified, most are aware that research malpractice and misconduct represents a major threat to their mission.

6.1 What should be done?

One way to resolve a problem is to try to address its major causes one by one.

A first path could therefore be to decrease the incentives to malpractices and misconduct, one of which is the pressure to publish for all physicians. Although encouraging research should be commended and is part of a strong academic curriculum, forcing all physicians, even those who have no interest in the research area, to produce research and to publish, is most likely counterproductive for the scientific knowledge and for the scientific community. According to Bastian et al, who recognise the difficulty of estimating these numbers, there were about 14 new RCTs published every day in the medical field in the 1980\textsuperscript{es}, and 75 RCTs and 11 systematic reviews
published every day in 2010, although “there are still only 24 hours in a day”...\textsuperscript{151} This enormous amount of scientific production ought to be better controlled. Academic promotion, as well as attribution of research funds, need to be based a little more on the quality of the articles published by the applicant and a little less on the quantity of published articles if we want to see a decrease in the number of useless research published.\textsuperscript{153} Also, when academic promotion and grant attribution will stop depending solely on the number of publications, especially in high impact factor journals, the temptation of adding guest-authors in publications,\textsuperscript{162} and even worse, the temptation to fabricate and/or falsify data may decrease as well.\textsuperscript{155} Unfortunately, research on the factors responsible for scientific misconduct remains sparse.\textsuperscript{163,164} It has been suggested that the lack of a clear definition of research misconduct, the lack of research integrity policy within an institution,\textsuperscript{99} and cultural characteristics restraining criticism were risk factors for fraud among young researchers.\textsuperscript{165} Also, if the number of publications is probably not a safe metric in order to evaluate a researcher, we have to admit that the evaluation of the scientific performance of both individuals and of institutions remains a challenge. Since year 2000, different metrics have been proposed in an attempt to quantify the scientific productions, such as for example Science citation index, Web of science, h-index or the more recent Mendeley and Altmetrics. However, these metrics are too often misused according to scientometricians who highlighted that reading a researcher’s work to judge his/her scientific quality is much more appropriate than relying on any given number.\textsuperscript{166} Future challenges for academic institutions will be to honestly re-think their structure and incentives for academic promotion,\textsuperscript{167} to find new innovative ways to evaluate their collaborators, and to replace the simple count of articles and summing up of impact factors with other, more reliable metrics.\textsuperscript{149,168,169}

A second path could be to implement structures to increase the quality of the research produced and published.\textsuperscript{10} This is likely to become necessary if medicine aims to remain a scientific discipline based on evidence. For example, physicians need to be better trained in research methodology and epidemiological principles throughout their medical curriculum. They need to be taught how to search for scientific evidence regarding a clinical question, where to look and how to critically appraise the published literature.\textsuperscript{165} They need to be aware of the importance of the science behind their everyday practice. All too often do I hear physicians dismissing the clinical trials as just merely serving to confirm what they already know from their clinical practice. A shift to a real scientific culture is needed among the medical community and this may take a long time if we consider where the medical culture and knowledge comes from. The Hippocratic Oath\textsuperscript{170} is an eloquent illustration of it’s “eminence based” foundations. A necessary step will require recognising that for today’s every-day practice, a physician needs to be able to search for literature and to critically appraise published evidence just as much as using a stethoscope. But also that performing good quality clinical trials requires very specific skills and that professional epidemiologist are needed in academic hospitals to help physician-researchers design, perform, report and publish high quality research. Also, if medicine aims to be a science, the philosophy
underlying the scientific process should be taught and highlighted repeatedly during the medical school. For example, physicians should clearly understand that although it may be quite pleasant to publish in *Science* or *Nature*, what is really important is to report what has been done during the research in such a way that readers can understand the procedures, to highlight any risk of bias or problem that may have distorted the reported results, including any potential conflicts of interests. There is no shame of not being able to replicate the results described in a previously published article. On the contrary, science needs to be replicated, and sometimes, because findings may happen just by chance, it is of major importance to be able to question previous findings.

A third path could be to improve the barriers to publication of poor quality and fraudulent research. As seen previously, there are up to now quite few existing barriers and most of them may be of limited use to avoid such publications. Much responsibilities relies on the shoulders of the ethics committees and of the reviewers and editors of scientific journals for spotting, correcting and sometimes for denouncing wrong doings, with very few counterparts for this difficult work. There are up to now no real incentives to become either a careful member of a research’s ethics committee or a rigorous peer-reviewer, the definition of peer-reviewer itself being quite vague. For example, peer-review is generally performed on a voluntary basis, without either pecuniary reward or academic recognition of the work performed. Making a thorough peer-review of a research article demands time and commitment, without counterpart. Open peer-review is starting to develop and may improve the motivation of peer-reviewers and the quality of their reviews, as their work no longer remains hidden. Some have suggested however that open peer-review made no difference in the quality, but did increase the number of peer-reviewers declining invitation. Unfortunately, research on peer-reviewing remain scarce. Some web-initiatives such as “Publon” are emerging to develop a better recognition of the work of peer-reviewers by publishing their peer-review history online, and rewarding “top peer-reviewers” with prizes. Their impact on the overall quality of peer-reviews remain to be demonstrated.

Finally, the last path could be to increase the post-publication verification process, which may identify errors or misconduct, lead to publication of corrections, and sometimes to retractions. Cautious readers sometimes take the time to write a letter to the Editor in Chief of a journal responsible for the publication of a flawed article after having spotted an error or a potential misconduct. We have illustrated this with a recent case in anaesthesiology, and shown how it could lead to identification of important frauds. Some web-platforms like “PubPeer” have also been recently developed to allow readers to comment and discuss published papers together; some discussion have led to the identification of cases of fraud. Other websites are specifically dedicated to highlighting multiple sorts of problems arising in scientific publication such as “Retraction Watch” that track retracted articles across all scientific journals. A system to better inform researchers regarding the retracted articles in their area of research is still needed, since it
has been shown that retracted articles are still cited after retraction.\textsuperscript{183} The positive side of this frightening reality, is that recognition of the existence of research misconduct and malpractice forces the scientific community to develop new procedures to try to avoid misconducts,\textsuperscript{98,184} User-based initiatives such as the "EvidenceLive Manifesto",\textsuperscript{185} or the "OpenTrials"\textsuperscript{186} initiatives are now regularly emerging, and will hopefully play an important role on the correction of the published literature, but also such as the "James Lind Alliance"\textsuperscript{187} to improve the relevance of research projects.

6.2 Where do systematic reviews fit into the picture?

There are several ways systematic reviews may be of interest in this context. One way systematic reviews may be useful is by helping to decrease the number of newly published trials. Every researcher should be asked, before he/she starts a new research project, to check whether a systematic review on the topic has already been performed and published. Depending on the quality of the systematic review published, one could either decide that the research is no longer needed since a definite answer to the research question has been reached, or, if this was not the case, the research agenda of the available systematic review could help the researcher to reformulate a research question that remains unclear and needs to be investigated further. If no systematic review is yet available on the topic, it could be a good idea to start by performing the missing systematic review, before starting yet a new research project. Only once a well-performed systematic review has clearly identified the need for further research should a new trial be started. This usage of systematic reviews, by researchers, but also by ethics committees,\textsuperscript{41} grant providers, or drug regulation agencies, could help to decrease the number of useless research published.

A second way systematic reviews could help in this context is by improving the methodological research knowledge of young physician-researchers. In fact, performing a systematic review during, or just after the medical curriculum, could be very formative. All of the post-graduate students or physicians (anaesthesiologists for most of them) I had the chance and the opportunity to accompany through the development of a systematic review, have attested of the profound impact that this experience has had on their understanding of the research methodology as well as the strengths and limitations of various study designs, including those of the systematic review itself (not discussed in the present dissertation). One needs to spend hours and days reading and critically appraising the published evidence on one given topic to really catch the reality of the science available. Many young researchers were surprised to realise the rather poor quality of the available evidence on which their everyday practice rely. Although this may sound negative, it is actually very important that every physician realises that reading one article is not enough to change their clinical practice, and that not everything that has been published necessarily reflects the truth. Additionally, the process of systematic reviewing\textsuperscript{188,189} as well as the standards of reporting systematic reviews\textsuperscript{190} are changing and improving regularly, making it a field of
continuous renewal. The implementation of “living systematic reviews”, the online and up-to-date summaries of available evidence, may represent the next step forward.\textsuperscript{191} Finally, it is of outmost importance that systematic reviews are taken for what they can offer, and not more. Although systematic reviews are vulnerable to the poor quality of research they rely on, they remain the best available source of evidence, as long as they are well performed, and acknowledge their limitations. Also, having read the entire scientific production answering a given question allows systematic reviewers to sometimes suspect or identify errors and misconducts. A systematic review is essentially a team work in which different experts from different fields are involved. One these experts should be a methodologist, epidemiologist or statistician which usually has a good capacity to detect potential errors in the reported data.\textsuperscript{192} Tools for suspecting the non-publication of small negative studies such as examining the plot of effect estimates in relation to the sample size of the trial have been recognised 20 years ago,\textsuperscript{122} and are used regularly by systematic reviewers. Other tools are less frequently used. For example it is quite difficult to fabricate totally “plausible” data and some statistical tools are now emerging to identify “doubtful” data such as a “very unlikely repartition of data in a randomised design”.\textsuperscript{103,156,193,194} The major issue that remains to be solved is what systematic reviewers ought to do when they come to suspect malpractices or misconduct. As discussed in chapter 5.5, most systematic reviewers do not seem to know what to do, nor do they seem willing to act as whistle-blowers. However, it should be the responsibility of all scientists to preserve the validity of published science.\textsuperscript{195,196} As the title of a recent editorial published by the Editors of PLoS Medicine stated “An unbiased scientific record should be everyone’s agenda”.\textsuperscript{197} It may be time for systematic reviewers to re-think their role in the matter. The Cochrane collaboration probably will need to play its role as a central actor in this context by creating a set of rules or recommendations as to what type of misconduct should be tracked, and define a clear code of conduct to help improve the validity of the accessible scientific record.\textsuperscript{198}
7 Conclusion

There is little doubt that a well-performed systematic review will always provide a higher level of evidence of the efficacy of an intervention compared to a single experimental trial. However, the accumulation of poor quality research and of research based on misconduct challenges its role. Although systematic reviews may not always provide definite (and correct) answers, they can be very useful to prevent the production of unnecessary trials and to guide researchers in their choices of relevant research questions, rigorous methodology and relevant outcomes to apply in new trials. Systematic reviews may also help to clean the published literature from errors and malpractices identified during the review process, and may help improve the methodological knowledge and appraisal skills of young researchers. Even though most of the answers to the difficult problem of research misconduct lie with academic structures, policies and regulations, and will require major transformation and paradigm shifts, systematic reviews may be seen as an interesting tool to support a better quality in the published medical literature.
8 Bibliography

19. Krumholz, H. M., Ross, J. S., Presler, A. H. & Egilman, D. S. What have we learnt from


157. Godlee, F., Smith, J. & Marcovitch, H. Wakefield’s article linking MMR vaccine and


