The poor reporting and use of statistical methods in orthopaedic papers has been widely discussed by both clinicians and statisticians. A detailed review of research published in general orthopaedic journals was undertaken to assess the quality of experimental design, statistical analysis and reporting. A representative sample of 100 papers was assessed for compliance to CONSORT and STROBE guidelines and the quality of the statistical reporting was assessed using a validated questionnaire. Overall compliance with CONSORT and STROBE guidelines in our study was 59% and 58% respectively, with very few papers fulfilling all criteria. In 37% of papers patient numbers were inadequately reported; 20% of papers introduced new statistical methods in the ‘results’ section not previously reported in the ‘methods’ section, and 23% of papers reported no measurement of error with the main outcome measure. Taken together, these issues indicate a general lack of statistical rigour and are consistent with similar reviews undertaken in a number of other scientific and clinical research disciplines. It is imperative that the orthopaedic research community strives to improve the quality of reporting; a failure to do so could seriously limit the development of future research.

Critical appraisal of research publications, for instance by peer review, is only possible if the methods and results of the studies are comprehensively and transparently reported. There is considerable anecdotal evidence of both poor reporting and use of statistical methods in orthopaedic papers by clinicians and statisticians. Petrie\(^1\) observed, from experience as a statistical reviewer, that the authors of a substantial proportion of papers submitted to a major orthopaedic journal had little understanding of statistical concepts: she reasoned that this may be attributed, in part, to the improved accessibility and ease of use of statistical software in recent years. Sophisticated analyses have now become much simpler to undertake with little or no formal statistical training. Evidence of poor statistical reporting is provided by Bhandari et al\(^2\) who highlighted the lack of blinding of outcome assessors and concealment of randomisation in randomised trials in orthopaedic research. Dulai et al\(^3\) reported that most randomised controlled trials (RCTs) in paediatric orthopaedic journals have substantial weaknesses in their methodology and in particular inadequate rigour and reporting of methods of randomisation, use of inappropriate or poorly described outcomes and inappropriate statistical analysis. Singh, Murphy and Bhandari\(^4\) also reported methodological deficiencies in RCTs of joint replacement, which compromise their quality. In response to such failures, Hartz and Marsh\(^5\) have provided guidance for observational studies in orthopaedic research.

In this study, we undertook a detailed review of research published in orthopaedic journals to assess the quality of experimental design, statistical analysis and reporting. A representative sample of 100 papers was collected using methods previously developed for this purpose.\(^6\) The primary aim was to highlight common errors and deficiencies in reporting, study design and statistical analysis in order to encourage improvement in the quality of published orthopaedic research. The focus of this paper is on the quality of reporting, which was assessed using published guidance for the reporting of clinical research (CONSORT and STROBE) and a questionnaire that assessed the quality of the statistical reporting in the sampled papers.

**Methods and Materials**

We sampled 100 papers from selected peer-reviewed general orthopaedic journals for a detailed review of reporting, design and statistical analysis.\(^7\) The following were used to identify the relevant papers for review: (1) studies reporting results of RCTs, case-control (CC) studies,
longitudinal cohort studies (L) and cross-sectional (CS) studies; (2) published research papers from seven general orthopaedic journals (Journal of Bone and Joint Surgery (American), Clinical Orthopaedics and Related Research, Journal of Bone and Joint Surgery (British), Acta Orthopae-
dica, Archives of Orthopaedic and Trauma Surgery, Inter-
national Orthopaedics and BMC Musculoskeletal Dis-
orders); (3) original research only, excluding trial proto-
col, reviews, meta-analyses, short reports, communications or letters; (4) published between 1 January 2005 and 1 March 2010 (study start date); (5) limited to no more than one paper from any single research group and to one paper only reporting a particular study or database; and (6) excluding papers published by research groups based at our own institutions.

The selected high-profile general orthopaedic journals included those with a wide range of impact factors; Journal of Bone and Joint Surgery (American) is ranked third and Archives of Orthopaedic and Trauma Surgery is ranked 32nd. The selection also included the first, third and fourth most cited orthopaedic journals from the Journal Citation Reports database at the Web of Knowledge. The selection of papers was limited to one paper from any single research group and one paper reporting results of a particular study or database to ensure that the survey results were not unduly influenced by the bad – or good – practice of one particularly productive group of researchers. Papers published by research groups at our own institutions were excluded to preclude the possibility of assessment bias.

A literature search was conducted in MEDLINE using the appropriate MeSH terms which corresponded to the selected study types (RCT, CC, L and CS) and their synonyms in order to identify relevant studies published in the selected journals. Their abstracts were reviewed by one of the authors (RH) who selected or rejected the papers on the basis of the exclusion criteria listed above. A total of 495 relevant papers were identified. A subsample of 120 papers was selected at random, using an allocation sequence from a computer random number generator, and their full texts were obtained and reviewed. Any paper that was not considered appropriate for the study was excluded: the final selection of 100 papers was made by random selection from the eligible papers (Fig. 1). The exclusions were reviewed and confirmed by both the statistical and

---

**Fig. 1**

Flow chart showing the process of paper selection and exclusion during the search process.
non-statistical members of the project team and a final list of papers agreed by the full team.

The form and structure of the questionnaire was agreed by the whole project team before the assessment process began. Each paper was read and a statistical questionnaire completed. After random ordering, odd-numbered papers were read by one statistician (NRP) and even-numbered papers by the other (CLP). Assessments were undertaken independently, after initial discussion and agreement on any issues that were considered to be problematic.

The basic form of the questionnaire was identical to one we had used previously, although the wording of a number of items was modified to suit orthopaedic research. A number of additional items were added to address issues that were considered to be particularly important for this study (e.g. details of blinding, reporting of error terms, etc.). The questionnaire covered the type of study, population under study, study design, outcome measures and the methods of statistical analysis and was designed to be completed relatively quickly, that is in roughly 30 to 45 minutes for a single paper. A subsample of 16 papers was randomly selected from the original sample of 100, and each of these was reviewed by both statisticians to determine the level of consistency between the two assessors.

**Standard reporting guidelines.** In addition to the statistical questionnaire, a clinically trained member of the study team (RH) assessed the RCTs using the Consolidated Standards of Reporting Trials (CONSORT) guidelines and the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines. These guidelines provide a standard assessment of the overall quality of reporting and highlight any areas of common weakness. A single percentage score was assigned to each study to indicate the proportion of correctly reported items and, thus, the compliance of the study with the relevant CONSORT or STROBE guidelines. A subsample of sixteen papers, randomly selected from the original sample of 100, was read independently by a second team member (NRP) and compliance scores were obtained in order to assess the reliability of these assessments.

**Statistical analysis.** The intraclass correlation coefficient (ICC) was used to assess the level of agreement between reviewers (RH and NRP) with regard to the compliance scores of the studies for CONSORT and STROBE, using the two-way random single measures test for absolute agreement. Agreement between the statistical reviewers (NRP and CLP) for individual items in the statistical questionnaire was estimated using the kappa statistic, a measure of chance-corrected agreement. The ICC and kappa statistic were calculated using the R statistical software and psy package for specialist psychometric functions, with a 95% confidence interval for the ICC determined after bootstrapping. A value greater than 0.7 for ICC and kappa statistics was considered sufficient to provide good evidence of agreement between reviewers. Data from the CONSORT, STROBE and statistical questionnaires were summarised as counts and percentage responses for each item in the respective questionnaire. Where appropriate, counts were compared between paper types using Fisher’s exact test and 95% confidence intervals (CI) calculated based on asymptotic normality.

**Results**

The ICC which assessed the agreement between the compliance scores made by the clinical and statistical reviewers was 0.99 (95% bootstrapped CI 0.80 to 1.00) and 0.95 (0.78 to 0.99) for CONSORT and STROBE respectively. Agreement values (kappa statistics) between the two statistical reviewers for each of the individual items in the questionnaire were in a range from 0.76 to 1.00, with a median of 1.00. The lowest level of agreement (0.76) was found for the item asking if the “purpose of the study was stated in the introduction”. This was the most subjective of the items in the questionnaire, so it is not surprising that this item had the lowest level of agreement. For most items tested there was perfect agreement; e.g. treatment group size and study type.

After detailed reading, five papers were found to report on a different type of study to that originally determined from the paper abstract. The final breakdown of study types reflected the relative prevalence in the literature; RCTs (n = 27), CC studies (n = 30), L cohort studies (n = 16) and CS studies (n = 27). On reading, each paper was classified into a small number of categories to indicate the anatomical site of the study (e.g. hip, foot and ankle, etc.). Table I shows the studies characterised by journal and anatomical site.

**CONSORT and STROBE compliance.** The overall compliance with the CONSORT and STROBE guidelines are shown by journal in Table I. Compliance with the CONSORT guidelines ranged from 11% (4 of 37) to 89% (33 of 37) for the 27 RCTs, with an overall compliance of 59% (95% CI 56 to 62). For the 73 case control, cross sectional and longitudinal studies, compliance with the STROBE guidelines ranged from 24% (8 of 33) to 79% (26 of 33), with an overall compliance of 58% (95% CI 56 to 60).

How the sample size was determined was reported in 50% (13 of 26) of RCTs and 42% (31 of 73) of observational studies. Details of the baseline characteristics of trial participants were reported in 81% (21 of 26) of RCTs and 59% (43 of 73) of observational studies. There was a discussion of the statistical methods in 88% (23 of 26) of RCTs and 79% (58 of 73) of observational studies. Identification of the study design in the title, using an explicit, commonly used term, is a common requirement for both CONSORT and STROBE. The study design was indicated by an appropriate term in the title or abstract in 78% (21 of 27) of RCTs and 53% (39 of 73) of observational studies. The source of funding was reported in 59% (16 of 27) of RCTs and 60% (44 of 73) of observational studies. Information about funding was often hidden at the end of
the text in small print with the acknowledgments, or was incomplete or vague. For instance, a number of studies simply stated that no funds were received from commercial companies, without any further information. Only one of the 27 RCTs in our sample reported a trial registered with a recognised database and stated the registration number. Both CONSORT and STROBE require that authors should report the number and type of settings and describe the care providers involved. In our study only 52% (14 of 27) of RCTs reported details of the setting and 63% (46 of 73) of observational studies reported details of the setting and approximate dates of recruitment and follow-up as required according to STROBE.

It is always important to consider and discuss the limitations of a study when reporting, and this is specifically recognised by both CONSORT and STROBE. Of the 100 papers sampled, 63% (17 of 27) of RCTs and 67% (49 of 73) of observational studies discussed the limitations of the study. Harm and unintended effects were reported in only 62% (16 of 26) of the RCTs.

Statistical questionnaire. Only 10% (10 of 100) of papers clearly stated the purpose of the study in the Introduction section of the paper. Gender was not reported in 11% of papers (11 of 100), making it impossible to identify clearly the population under study. Patient age was reported in 93% (93 of 100) of papers and weight or body mass index (BMI) in 40% of papers (40 of 100). In 6% (6 of 100) of papers it was unclear whether reported analyses were comparing, for instance, responses between groups of patients (between subject analysis) or changes in response over time for each patient (within subject analysis).

Patient numbers were either stated exactly (100 patients were included in this study), as an estimate (between 100 and 110 patients were eligible for recruitment) or were not reported. Table II shows that only 57% (57 of 100) of papers reported exact numbers of patients in both the Methods and Results sections. A sizeable proportion of papers (37%; 37 of 100) reported numbers in only one of these two sections. Of the 57 papers that reported exact numbers throughout, seven (12%) stated larger numbers in the Methods than the Results section and two (4%) reported smaller numbers in the Methods than the Results section. Numbers in the Methods may be larger than those reported in the Results due to, for instance, patient withdrawal from a trial, death or other loss to follow-up, but it is unclear why the number of patients reported in the Results should ever be larger than that reported in the Methods.

Statistical methods were described in the Methods section in 96% (96 of 100) of papers, but in 20% (20 of 100) of papers statistical methods not previously reported in the Methods section were reported in the Results section. Unknown or uncertain statistical analyses, for instance p-values stated with no explanation of the statistical test used, were reported in 4% (4 of 100) of papers. When reporting analyses, point estimates such as means should always be presented with errors that quantify the variability or precision of the measure, e.g. standard deviation (SD), standard error or CI. Identifiable error estimates were reported in 77% (77 of 100) of papers: the remaining 23% (23 of 100) of papers reported exact numbers of patients in both the Methods and Results sections. A sizeable proportion of papers (37%; 37 of 100) reported numbers in only one of these two sections. Of the 57 papers that reported exact numbers throughout, seven (12%) stated larger numbers in the Methods than the Results section and two (4%) reported smaller numbers in the Methods than the Results section. Numbers in the Methods may be larger than those reported in the Results due to, for instance, patient withdrawal from a trial, death or other loss to follow-up, but it is unclear why the number of patients reported in the Results should ever be larger than that reported in the Methods.

Statistical methods were described in the Methods section in 96% (96 of 100) of papers, but in 20% (20 of 100) of papers statistical methods not previously reported in the Methods section were reported in the Results section. Unknown or uncertain statistical analyses, for instance p-values stated with no explanation of the statistical test used, were reported in 4% (4 of 100) of papers. When reporting analyses, point estimates such as means should always be presented with errors that quantify the variability or precision of the measure, e.g. standard deviation (SD), standard error or CI. Identifiable error estimates were reported in 77% (77 of 100) of papers: the remaining 23% (23 of 100)
reporting either no error measurements (19%; 19 of 100) or error measurements that were not identifiable (4%; 4 of 100). Also, 30% (30 of 100) of papers presented errors in the format ‘mean +/- SD’. This format for summarising outcomes is generally disliked by statisticians, with the simpler ‘mean (SD)’ being recommended as a less ambiguous way of expressing the variability in the data. A full summary of the results for each item in the statistical questionnaire is available on request from the corresponding author.

Discussion

The work reported here used standard guidelines for the reporting of studies (CONSORT and STROBE) as well as a questionnaire to review the quality of reporting in a random sample of 100 papers from the orthopaedic literature. A carefully designed sampling strategy ensured that the papers were representative of the target literature. There was good evidence of agreement (ICC and kappa statistics) between reviewers for a selected subsample of papers for both CONSORT and STROBE, and for the statistical questionnaire.

Overall compliance with CONSORT and STROBE guidelines in our study was 59% and 58% respectively; very few papers complied with all criteria. Similar studies in other specialties have usually limited their attention to RCTs and CONSORT only. For example, compliance in obstetric anaesthesia journals ranges from 36% to 100% with a median score of 65%: equivalent median scores for our survey were 65% for CONSORT and 57% for STROBE. It may not always be possible for trials involving surgical interventions to meet all criteria, such as performing sham interventions or blinding patients. However, the poor rate of compliance of these papers to the minimum standards of reporting is a major concern. There is good evidence that CONSORT has improved the quality of reporting trials. As such, journal editors, authors and reviewers alike have a responsibility to insist on adherence to these standards. It is encouraging that a number of orthopaedic journals have now adopted both CONSORT and STROBE, in line with most of the major general medical journals.

It is important that authors describe in detail the role of the funding bodies and their level of involvement and influence on design, conduct and analysis of the data. In our study, the source of funding was reported in 59% of RCTs and 60% of observational studies. Previous papers which reported the results of trauma and orthopaedic RCTs found that only 55% made any declaration of the funding source. Although troubling, this level of poor reporting is not uncommon. In a survey of 370 drug trials, 29% failed to report sources of funding. In another survey of PubMed indexed randomised trials published in December 2000, the source of funding was reported for 66% of the 519 trials.

Information about the settings and locations of a study is crucial for assessing applicability and generalisability. Authors should report the number and type of settings and describe the care providers involved. In particular, studies should clarify whether a trial was carried out in one or several centres, and provide enough information so that the reader can judge whether the results of the trial are relevant to their own setting. The low proportion of RCTs (52%) and observational studies (63%) providing full details of the setting show that this was not the case for the papers sampled here.

Identification and discussion of the weaknesses of a study are particularly important and are needed for a balanced interpretation of the results. The difference between statistical significance and clinical importance needs to be considered and, for RCTs, the discussion should avoid the common error of interpreting a non-significant result as indicating equivalence of the interventions. The fact that only 63% of RCTs and 67% of observational studies discussed the limitations of the study was surprising. Similarly, information about the harm as well as the benefits of interventions is stipulated by the CONSORT guidelines. However, details of harm and unintended effects were reported in just 62% of the RCTs in our sample.

Trial registration is stipulated by the CONSORT guidelines and the World Health Organization states that the registration of all interventional trials is a scientific, ethical and moral responsibility. In 2004, the International Committee of Medical Journal Editors (ICMJE) implemented a policy change to state that they would only consider trials for publication if they had been registered before the enrolment of the first participant. This resulted in a dramatic increase in the number of trials being registered, and an associated improvement in the quality of reporting for registered trials. However, only one of the 27 RCTs in our sample reported a trial registered with a recognised database and stated the registration number.

The poor compliance of these papers to STROBE and CONSORT guidelines was mirrored in the overall quality of the statistical reporting. In 6% of papers it was not clear whether comparisons were within or between subjects; 37% of papers inadequately reported patient numbers; 20% of papers introduced new statistical methods in the Results section not previously reported in the Methods section and 23% of papers stated no error measurements with the main outcome measures. Many of these issues may seem trivial in themselves, but taken together they indicate a general lack of statistical rigour in reporting. Most of the problems identified could have been rectified by appropriate statistical support at an early stage in the study design, or at least should have been flagged-up by routine statistical review of the manuscript on submission to the journal.

This study has some limitations. A full systematic review of the orthopaedic literature was not undertaken because the resources available were limited. However, a rigorous framework was used for randomly selecting papers from a representative sample of journals to ensure that the study provided a realistic overview of the state of reporting in...
orthopaedic research. Also, for pragmatic reasons, both of the statistical authors were not able to read each paper, due simply to time constraints. However, this is not a major concern as the assessment of agreement for a sub-sample of papers suggested a high degree of consistency between assessors.

The problems with reporting identified here are consistent with similar reviews undertaken in a number of other scientific and clinical research disciplines, which also report major deficiencies in design, analysis and reporting.6,22,23 The orthopaedic research community relies on published trials and other experimental studies being appropriately designed, analysed and accurately reported. It is therefore imperative that we strive to improve the quality of reporting; a failure to do so could seriously limit the development of future research.

Supplementary material

Tables detailing the statistical questionnaire, CONSORT compliance for the randomised controlled trials, and STROBE compliance for the observational studies are available with the electronic version of this paper on our website www.jbjs.org.uk

Listen live

Listen to the abstract of this article at www.jbjs.org.uk/interactive/audio

No benefits have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

References