American Journal of Epidemiology

Vol. 161, No. 3
Printed in U.S.A.
DOI: 10.1093/aje/kwi042

PRACTICE OF EPIDEMIOLOGY

Quality of Reporting of Observational Longitudinal Research

Leigh Tooth[1], Robert Ware[1], Chris Bain[1], David M. Purdie[1,2], and Annette Dobson[1]

[1] School of Population Health, The University of Queensland, Brisbane, Queensland, Australia.
[2] Population Studies and Human Genetics Division, Queensland Institute of Medical Research, Brisbane, Queensland, Australia.

Received for publication July 9, 2004; accepted for publication August 31, 2004.

Observational longitudinal research is particularly useful for assessing etiology and prognosis and for providing evidence for clinical decision making. However, there are no structured reporting requirements for studies of this design to assist authors, editors, and readers. The authors developed and tested a checklist of criteria related to threats to the internal and external validity of observational longitudinal studies. The checklist criteria concerned recruitment, data collection, biases, and data analysis and descriptive issues relevant to study rationale, study population, and generalizability. Two raters independently assessed 49 randomly selected articles describing stroke research published from 1999 to 2003 in six journals: *American Journal of Epidemiology*, *Journal of Epidemiology and Community Health*, *Stroke*, *Annals of Neurology*, *Archives of Physical Medicine and Rehabilitation*, and *American Journal of Physical Medicine and Rehabilitation*. On average, 17 of the 33 checklist criteria were reported. Criteria describing the study design were better reported than those related to internal validity. No relation was found between study type (etiologic or prognostic) or word count and quality of reporting. A flow diagram for summarizing participant flow through a study was developed. Editors and authors should consider using a checklist and flow diagram when reporting on observational longitudinal research.

epidemiologic factors; longitudinal studies

Abbreviations: CONSORT, Consolidated Standards of Reporting Trials; SD, standard deviation.

Reporting requirements for clinical trials have improved substantially since the 1960s, when researchers first identified a lack of rigor (1–3). The movement toward standardized reporting has arisen from the recognition that inadequate reporting, for example, of concealed allocation, can lead to biased interpretation (4). The tangible outcome of this improvement is the revised Consolidated Standards of Reporting Trials (CONSORT) statement, comprising a 22-item checklist and flow diagram (5). CONSORT has been adopted by over 150 journals worldwide (6), and its use has been linked with improved quality of reporting of clinical trials, although inadequacies persist (7–9). Since CONSORT, other statements of reporting requirements for nonrandomized interventions (10), meta-analyses (11, 12), and diagnostic tests (13) have appeared. As yet, however, no such equivalent standards for reporting observational longitudinal studies are known to exist. The strength of this design, particularly for assessing etiology and prognosis, is increasingly being recognized (14–17), as is the value of the evidence for clinical decision making (18–21).

In the absence of standard reporting guidelines, authors may refer to theoretical papers and texts describing observational longitudinal research designs (22–24). Although some of these sources provide comprehensive coverage of aspects of observational longitudinal research on which internal and external validity of results depend, others are brief. A few authors have developed checklists with which to assess the quality of reporting of articles, including observational longitudinal research (10, 25–28). These checklists differ in their coverage of elements relevant to the design of observational longitudinal research. The majority are brief or nonspecific and focus on quality judgments. The most comprehensive of these is the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) state-

Correspondence to Dr. Leigh Tooth, School of Population Health, University of Queensland, Herston Medical School, Herston, Queensland, 4006 Australia (e-mail: L.tooth@sph.uq.edu.au).

ment (10), which is very detailed and places a particular emphasis on interventions. It provides a detailed assessment of the quality of these designs and has suggestions for better reporting. However, none of the checklists offers a simple or straightforward set of guidelines for how observational longitudinal studies should be reported. Adequate reporting is the only means by which proper interpretation can occur (20). The success of CONSORT illustrates the benefits to be gained from improved communication between authors, editors, and readers about research design fundamentals.

The aim of this study was to identify desirable elements in the reporting of observational longitudinal research, construct a CONSORT-style checklist and flow diagram, and test the checklist against published observational longitudinal research. Like other authors (3, 27), we focused on the adequacy of reporting (i.e., whether or not an aspect was reported) and did not attempt to assess quality per se. The secondary focus was to explore the likely value to editors and authors of developing a checklist and flow diagram, covering desirable reporting elements, that would help readers evaluate observational longitudinal research.

MATERIALS AND METHODS

Development of guidelines

We examined the literature on reporting of observational longitudinal research by using the MEDLINE (National Library of Medicine, Bethesda, Maryland), PSYCHLIT (American Psychological Association, Washington, DC), and CINALH (Cinahl Information Systems, Glendale, California) online databases and by hand searching. Search terms included observational, longitudinal, prospective, follow-up, cohort, and outcomes. The literature retrieved described threats to the internal and external validity of longitudinal research and epidemiologic methods in general (e.g., Grimes and Schulz (16), McKee et al. (18), the Epidemiology Work Group of the Interagency Regulatory Liaison Group (22), Wolfe (23), Hartz and Marsh (24), Kleinbaum et al. (29), Greenland (30), Zapf et al. (31), Wolfe et al. (32), Zaccai (33), and Grimes et al. (34)). Several authors had developed their own checklists to assess reporting in epidemiologic studies, and these were reviewed (10, 25–28). Published checklists for reporting randomized (5) and nonrandomized trials (10), meta-analyses (11, 12), and reports on diagnostic accuracy (13) were also reviewed. Additionally, we also examined textbooks on epidemiology (e.g., Rothman and Greenland (35) and Hennekins and Buring (36)).

A draft outline of essential elements related to threats to the internal validity of observational longitudinal research was created. A working group of nine epidemiologists, biostatisticians, and social scientists, with a wide range of qualifications, experience, and clinical interests, contributed and revised checklist criteria. For each essential element identified (e.g., selection bias), the most important criteria (descriptors) to describe an observational longitudinal study were identified (e.g., sampling frame, consent rates, loss to follow-up, item nonresponse). Through this iterative revision process, other criteria fundamental to describing observational longitudinal research adequately (e.g., setting) and to considering generalizability were added to the checklist. Criteria were to be scored as reported (yes), not reported (no), or not applicable to report. To score "yes," each criterion must be reported in enough detail to allow the reader to judge that the definition had been met. If inadequate information about a criterion was reported, it was scored "no." If authors referred readers to another publication for specific details about the study methods (e.g., sampling or eligibility), the criterion was scored "no."

The draft checklist was piloted by the first two authors (L. T. and R. W.), who independently rated 10 articles describing observational longitudinal research (defined as studies in which any designated group of persons was followed or traced over a period of time) (37). Following the pilot study, the criteria were reviewed and were modified by the working group. Once the final checklist was agreed upon, it was tested on a random selection of articles describing observational longitudinal research. The clinical area of stroke was chosen as an example because it is the current field of interest of the first author. None of the other authors or members of the working group had substantive experience in stroke research. Six journals publishing epidemiology, clinical, and rehabilitation stroke research, with a range of impact factors (from 0.9 to 8.6), were chosen: *American Journal of Epidemiology*, *Journal of Epidemiology and Community Health*, *Stroke*, *Annals of Neurology*, *Archives of Physical Medicine and Rehabilitation*, and *American Journal of Physical Medicine and Rehabilitation*.

Ten articles reporting observational longitudinal research were randomly sampled from each journal. The sampling frame was every volume of the six journals published between June 1999 and June 2003 inclusive. Ten randomly generated volume/issue "pairs" (e.g., issue 3, 2002) were produced for each journal. Potentially eligible articles were identified from words such as "longitudinal," "follow-up," "outcomes," "prospective," or "observational" appearing in the title or abstract. Content eligibility was assessed by the presence of any of the words "stroke," or "cerebrovascular accident," or "CVA," or "acquired brain injury," or "infarct" coupled with a structure or hemisphere of the brain; or words illustrative of stroke symptoms, for example, "hemiplegia," "hemiparesis," or "neglect." Exclusion criteria were words indicating that the study was randomized; an intervention; a case series; a case-control, cross-sectional, or retrospective study; or a systematic review. Studies of animals were also excluded. When more than one eligible article was identified in a particular volume/issue pair for a selected journal, all were numbered and one selected randomly. When a volume/issue pair had no eligible articles, a new volume/issue pair was randomly generated for the same journal. The *American Journal of Epidemiology* and the *Journal of Epidemiology and Community Health* had only three and six eligible articles, respectively, within the sampling frame, so all were included. None of the authors or the members of the working group was an author of any of the sampled publications.

Of the 49 articles selected, six were published from June to December 1999, 11 during 2000, 10 during 2001, and 11 each during 2002 and from January to June 2003. The article list is available at the following website: http://

www.sph.uq.edu.au/hisdu/bias_refs.html. Each article was independently rated with the checklist by the first two authors, who then compared ratings and resolved disagreements by consensus. When disagreements could not be resolved, a third independent rater made the final judgment. Besides the rating of each article with the checklist, it was noted whether the study was primarily etiologic ($n = 20$), prognostic ($n = 25$), or both ($n = 4$). The text word count of each article was also estimated. The working group also drafted a summary flow diagram to represent the essential elements of participant recruitment and follow-up in observational longitudinal studies.

Statistical analysis

Descriptive statistics were computed for each checklist criterion by type of study (etiologic or prognostic), journal, and word count. Agreement between the two raters on the 33 criteria was summarized by percentage agreement, presented here by the median and quartiles. For each article, the number of criteria reported was divided by the number of relevant criteria to give a score reflecting the proportion of relevant or applicable criteria reported. For example, if 12 criteria were reported when 33 were applicable, the proportion was 0.36; if 12 criteria were reported when 31 were applicable, the proportion was 0.39. The comparison between type of study (etiologic or prognostic) and proportion of criteria reported was analyzed by using an independent-samples *t* test. The association between estimated word count and the proportion of criteria reported was analyzed by using Spearman's correlation coefficient. Analyses were performed with SPSS software (version 11.5; SPSS Inc., Chicago, Illinois).

RESULTS

Development of the checklist

The final checklist comprised 33 criteria (table 1). The definitions used for the criteria and their sources are also included in table 1. The criteria reflect design and interpretation aspects covering the study rationale and population, recruitment, measurement and biases, data analysis, and generalizability of the results. The criteria represent two principal categories: 1) aspects that could possibly influence effect estimates and 2) more descriptive or contextual elements. Not all criteria were deemed applicable to all studies. For example, in some epidemiologic studies, the investigators do not have access to data on the nonconsenting members of the target population and cannot then compare them with consenters.

Application of the checklist to the 49 articles

For the two independent raters, the median percentage agreement was 75 percent (quartiles 62 percent–93 percent). The criteria that the raters had to discuss most often were number of participants at each stage, reliability of measurement methods, validity of measurement methods, reasons for loss to follow-up at each stage, missing data items at each stage, and absolute effect sizes. The raters resolved most coding discrepancies by consensus. A third independent rater was required to make the final decision about reliability of measurement methods in three articles and validity of measurement methods in one article.

Across the 49 articles, the mean proportion of applicable criteria reported was 0.51 (standard deviation (SD), 0.15; range, 0.12–0.82). The association between type of study (etiologic or prognostic) and proportion of criteria reported was not statistically significant ($t_{(43\ df)} = 0.31$, $p = 0.76$, two sided; studies with both an etiologic and prognostic focus, $n = 4$, were not included). When analyzed by journal type, the mean proportions of applicable criteria reported by the journals were 0.66 (SD, 0.03) for the *American Journal of Epidemiology* (impact factor = 4.2), 0.57 (SD, 0.11) for the *Journal of Epidemiology and Community Health* (impact factor = 2.1), 0.54 (SD, 0.13) for *Archives of Physical Medicine and Rehabilitation* (impact factor = 1.3), 0.49 (SD, 0.13) for *Stroke* (impact factor = 5.1), 0.46 (SD, 0.13) for the *American Journal of Physical Medicine and Rehabilitation* (impact factor = 0.9), and 0.46 (SD, 0.19) for *Annals of Neurology* (impact factor = 8.6). We found no relation between word count and proportion of checklist criteria reported (Spearman's correlation coefficient = 0.12, $p = 0.41$, two sided).

Table 2 shows the total number of articles that reported each of the 33 criteria overall and by type of study. The table also shows the total number (and percentage) of articles where it was applicable to report each of the criteria. Eleven articles had one or more criteria that were not applicable to report. Table 2 shows that "reasons for loss to follow-up at each stage," accounting for "loss to follow-up in the analysis," and "missing data in the analysis" were the criteria to which "not applicable" most often applied.

In total, 16 articles (nine etiologic, seven prognostic) referred the reader to another publication for methodological details. In 13 articles, this referral was accomplished directly by using wording such as "full details are reported elsewhere"; three articles were less direct, citing a reference to a previous publication that used the same data.

The best reporting was for criteria describing the study rationale and population as well as how data were collected and analyzed (each criterion reported in 45 or more articles). Qualitative and quantitative assessments of bias (30–35 articles) and confounders (38 articles) were also generally well reported. The most poorly reported criteria (reported in fewer than 10 articles each) were justification for the numbers in the study (e.g., in terms of power to detect effects), reasons for not meeting eligibility criteria, numbers consenting/not consenting, reasons for nonconsent, comparison of consenters with nonconsenters, and accounting for missing data items or loss to follow-up in analyses. Also notable was the general lack of reporting of measures of absolute effects, even though it is regularly described in epidemiology textbooks as a particular strength of observational longitudinal studies.

Development of the flow diagram

As a result of developing the checklist and rating the articles, we produced a flow diagram, modeled on CONSORT (5), to help clarify the numerical history of an observational longitudinal study (figure 1). It records the numbers, and

TABLE 1. The checklist criteria, and their definitions, used to rate the articles included in the study*

Criterion	Definition
1. Are the objectives or hypotheses of the study stated?	Self-explanatory.
2. Is the target population defined?	The group of persons toward whom inferences are directed. Sometimes the population from which a study group is drawn.
3. Is the sampling frame defined?	The list of units from which the study population will be drawn. Ideally, the sampling frame would be identical to the target population, but it is not always possible.
4. Is the study population defined?	The group selected for investigation.
5. Are the study setting (venues) and/or geographic location stated?	Comment required about location of research. Could include name of center, town, or district.
6. Are the dates between which the study was conducted stated or implicit?	Self-explanatory.
7. Are eligibility criteria stated?	The words "eligibility criteria" or equivalent are needed, unless the entire population is the study population.
8. Are issues of "selection in" to the study mentioned?†	Any aspect of recruitment or setting that results in the selective choice of participants (e.g., gender or health status influenced recruitment).
9. Is the number of participants justified?	Justification of number of subjects needed to detect anticipated effects. Evidence that power calculations were considered and/or conducted.
10. Are numbers meeting and not meeting the eligibility criteria stated?	Quantitative statement of numbers.
11. For those not eligible, are the reasons why stated?	Broad mention of the major reasons.
12. Are the numbers of people who did/did not consent to participate stated?	Quantitative statement of numbers.
13. Are the reasons that people refused to consent stated?	Broad mention of the major reasons.
14. Were consenters compared with nonconsenters?	Quantitative comparison of the different groups.
15. Was the number of participants at the beginning of the study stated?	Total number of participants (after screening for eligibility and consent) included in the first stage of data collection.
16. Were methods of data collection stated?	Descriptions of tools (e.g., surveys, physical examinations) and processes (e.g., face-to-face, telephone).
17. Was the reliability (repeatability) of measurement methods mentioned?	Evidence of reproducibility of the tools used.
18. Was the validity (against a "gold standard") of measurement methods mentioned?	Evidence that the validity was examined against, or discussed in relation to, a gold standard.
19. Were any confounders mentioned?	Confounders were defined as a variable that can cause or prevent the outcome of interest, is not an intermediate variable, and is associated with the factors under investigation.
20. Was the number of participants at each stage/wave specified?	Quantitative statement of numbers at each follow-up point.
21. Were reasons for loss to follow-up quantified?	Broad mention and quantification of the major reasons.
22. Was the missingness of data items at each wave mentioned?	Differences in numbers of data points (indicating missing data items) explained.
23. Was the type of analyses conducted stated?	Specific statistical methods mentioned by name.
24. Were "longitudinal" analysis methods stated?	Longitudinal analyses were defined as those assessing change in outcome over two or more time points and that take into account the fact that the observations are likely to be correlated.
25. Were absolute effect sizes reported?	Absolute effect was defined as the outcome of an exposure expressed, for example, as the difference between rates, proportions, or means, as opposed to the ratios of these measures.
26. Were relative effect sizes reported?	Relative effects were defined as a ratio of rates, proportions, or other measures of an effect.
27. Was loss to follow-up taken into account in the analysis?	Specific mention of adjusting for, or stratifying by, loss to follow-up.
28. Were confounders accounted for in analyses?	Specific mention of adjusting for, or stratifying by, confounders.
29. Were missing data accounted for in the analyses?	Specific mention of adjusting for, or stratifying by, or imputation of missing data items.
30. Was the impact of biases assessed qualitatively?	Specific mention of bias affecting results, but magnitude not quantified.
31. Was the impact of biases estimated quantitatively?	Specific mention of numerical magnitude of bias.
32. Did authors relate results back to a target population?	A study is generalizable if it can produce unbiased inferences regarding a target population (beyond the subjects in the study). Discussion could include that generalizability is not possible.
33. Was there any other discussion of generalizability?	Discussion of generalizability beyond the target population.

* Sources for definitions: Rothman and Greenland (35), Last (37), Twisk (41).
† Represents selection bias at the beginning of a study. Other selection biases (i.e., loss to follow-up, missing data items) are dealt with by other checklist criteria.

reasons for, eligibility, consent, participation in each wave, and attrition. These main elements were chosen because they provide information at a glance on probable selection-driven threats to internal and external validity.

TABLE 2. Numbers of articles in each journal reporting "yes" for each criterion/total number of articles where it was applicable to report that criterion, for all articles and by type of study

Criterion	All articles (n = 49)		Type of study*			
			Etiology (n = 20)		Outcomes (n = 25)	
	No.	%	No.	%	No.	%
1. Objectives/hypotheses	46/49	94	18/20	90	24/25	96
2. Target population	33/49	67	14/20	70	16/25	64
3. Sampling frame	41/49	84	16/20	80	21/25	84
4. Study population	45/49	92	19/20	95	22/25	88
5. Study setting/geographic location	39/49	79	16/20	80	19/25	76
6. Dates	35/49	71	14/20	70	19/25	76
7. Eligibility criteria	32/49	65	11/20	55	18/25	72
8. Selection-in biases	14/49	28	5/20	25	7/25	28
9. Number at beginning justified	0/49	0	0/20	0	0/25	0
10. Numbers meeting eligibility criteria	13/49	26	6/20	30	7/25	28
11. Reasons for not meeting eligibility criteria	6/49	12	4/20	20	2/25	8
12. Numbers consenting	9/49	18	2/20	10	6/25	24
13. Reasons for not consenting†	1/47	2	0/20	0	1/23	4
14. Comparison of consenters with nonconsenters†	1/47	2	1/20	5	0/23	0
15. Number of participants at the beginning	49/49	100	20/20	100	25/25	100
16. Method of data collection	47/49	96	18/20	90	25/25	100
17. Reliability of measurement methods	20/49	41	6/20	30	12/25	48
18. Validity of measurement methods	19/49	39	8/20	40	10/25	40
19. Confounders	38/49	77	17/20	85	17/25	68
20. Number of participants at each stage	25/49	51	8/20	40	14/25	56
21. Reasons for loss to follow-up at each stage‡	22/42	52	9/18	50	9/21	43
22. Missing data items at each stage	19/49	39	6/20	30	11/25	44
23. Type of analyses§	47/48	98	20/20	100	25/25	100
24. Longitudinal methods§	36/48	75	16/20	80	19/25	76
25. Absolute effect sizes§	3/48	6	3/20	15	0/25	0
26. Relative effect sizes§	18/48	37	13/20	65	5/25	20
27. Loss to follow-up in the analysis‡,§	2/41	5	1/18	5	1/21	5
28. Confounders in analysis§	29/48	60	17/20	85	11/25	44
29. Missing data in the analysis§,¶	2/42	5	0/19	0	2/22	9
30. Biases assessed qualitatively	35/49	71	13/20	65	19/25	76
31. Biases estimated quantitatively	30/49	61	17/20	85	11/25	44
32. Relate results to target population	37/49	75	15/20	75	19/25	76
33. Other discussion of generalizability	26/49	53	9/20	45	16/25	64

* Studies with both types (n = 4) were not included.
† Not applicable to report for two articles because every patient consented.
‡ Not applicable to report for seven articles because no loss to follow-up occurred.
§ Not applicable to report for one article because only descriptive summary measures were presented, and no analysis was conducted.
¶ Not applicable to report for six articles because no data were missing.

DISCUSSION

More than 20 years ago, DerSimonian et al. wrote in relation to clinical trials that "although all may not agree on our specific list of items, editors could greatly improve the reporting … by providing authors with a list of items that they expect to be strictly reported" (3, p. 1336). They pointed out that while weakness in design may occur for good reason, weakness in reporting should not occur. Their statements apply just as cogently to observational longitudinal research, and use of a checklist such as ours may be useful to help prevent weak reporting.

We have shown variable reporting of some of the major threats to the internal and external validity of observational longitudinal studies. In the articles sampled, on average

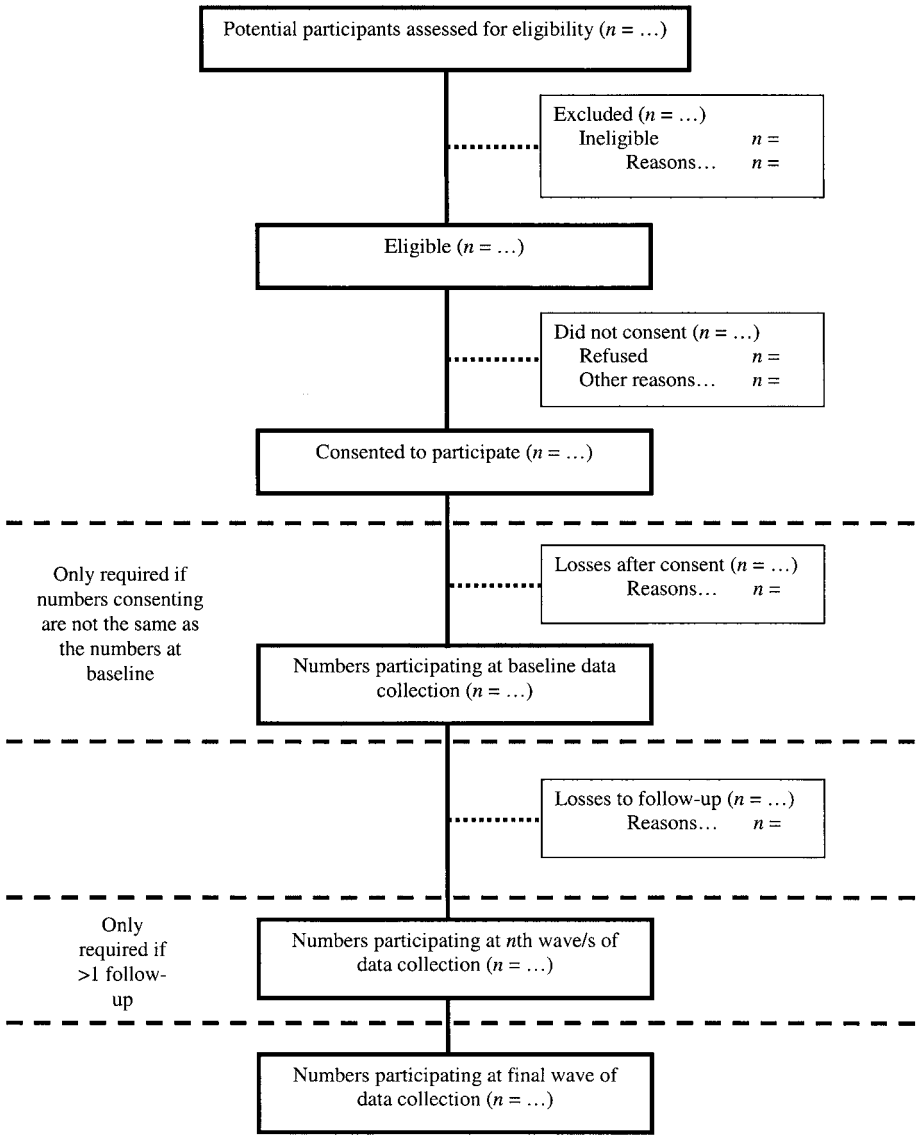

FIGURE 1. Flow diagram created to clarify participation in observational longitudinal research.

about half of the 33 checklist criteria were reported, with no differences found between study type or by word count. The criteria in the checklist representing selection bias were the least frequently reported overall, although issues of measurement quality were also neglected, with fewer than half of the articles discussing either reliability or validity. These findings are concerning because if observational longitudinal studies are to be accepted as valuable sources of evidence, complete reporting is required.

Aspects of recruitment, particularly the proportion of sampled subjects meeting the eligibility criteria and then consenting to participate, were poorly reported. In addition, the reasons that people did not consent, and comparisons of consenters with nonconsenters in terms of baseline demographic or clinical features, were also typically not reported. These aspects of selection bias are potentially important; if consenters differ from nonconsenters, the study findings may be affected. Dunn et al. (38) recently showed nonconsent in five large epidemiologic studies to be about 30 percent and illustrated how nonconsenters and nonresponders can account for 30–60 percent of the original sample. They recommend that researchers plan a priori their sample sizes to account for potential losses and consider the biases likely to be associated with nonconsent and dropout.

Although the numbers of participants at each stage of a study were recorded in half the articles, accounting for loss to follow-up and missing data items in the analyses were rarely reported. Data missing not at random can be a source of bias affecting internal validity and can also influence estimates of absolute prevalence or incidence (39, 40). In this study, we assessed how missing data were handled by whether the articles described imputation, weighting, or

sensitivity analyses. It is acknowledged that while authors may not statistically account for missing data in this way, they may postulate on the likely impact of missing data on results. When authors did so, it was captured under criterion 31 of the checklist: "Was the impact of biases estimated quantitatively?" Approximately 60 percent of articles acknowledged the possible quantitative impact of various biases, illustrating a general awareness by authors, or determination by editors, of the necessity for doing so. Methods for dealing with missing data in observational longitudinal research range from simple analysis of between-group differences to complex imputation techniques (41). Although debate exists about the benefits of using such imputation methods, it is at least desirable to determine the pattern of missingness, how ignorable or informative the missing data are, and the potential impact that imputation or other approaches may have on the final estimates (40).

None of the 49 articles included any justification for the sample size. An issue for many longitudinal observational studies is lack of statistical power or precision to determine real differences until sufficient follow-up time has passed to accumulate enough outcomes (42). Although the appropriateness of calculating statistical power for these research designs has been questioned (41), a priori consideration of the precision of a longitudinal study to accurately quantify the difference between effects of exposures on an outcome is desirable (35, 38).

Absolute effect sizes, defined in this study as the difference in rates of disease between groups defined by an exposure, for example, attributable risk, were also infrequently reported. Inclusion of this criterion was strongly debated by the working group because it is not relevant for all observational longitudinal studies. However, absolute effect estimates are a useful measure of association in epidemiologic research (39) and are an underutilized strength of observational longitudinal studies. In the checklist, absolute effects can be seen primarily as a descriptive criterion rather than as an element representing threats to internal validity.

About 40 percent of the articles reported the reliability and validity of instruments used. In a study of reporting of psychometric qualities of measures in 171 articles describing rehabilitation studies, Dijkers et al. (43) also found poor reporting, with reliability and validity mentioned in only 20 percent and 7 percent of articles, respectively. Having reliable and valid instruments is one of the best ways of reducing measurement bias in epidemiologic research. Requiring authors to report these psychometric properties may improve the quality of the instruments used, and the confidence with which conclusions can be drawn from the results. Obviously, this requirement is unrealistic for every measure in a long list of variables, but it is desirable to have some assessment of measurement quality for the core variables, including confounders, in a particular analysis.

Only four criteria were universally reported in the articles: the study objectives, the study population, the number of participants at the beginning, and the method of data collection. Criteria about confounding, and actions to account for confounding in the analysis, were also generally well reported (in more than 60 percent of the articles). This issue is important because confounding is one of the major limitations of nonrandomized designs such as observational longitudinal studies, and adjustment in the analysis is essential for identifying true effects.

Despite the variable reporting of actions taken to reduce bias, chance, and confounding, three quarters of the articles discussed generalizability of the results to the target population. In some cases, authors acknowledged caveats to generalizability because of limitations such as selection bias. However, it is important to recognize that generalizability should be considered only once assumptions of internal validity are satisfied.

We have shown a need for improved reporting of observational longitudinal research, through application of a reasonable set of criteria and a flow diagram. Even though the clinical example used in this study was stroke, the checklist and flow diagram are independent of topic and so are directly applicable to other fields. If authors are required to report criteria such as those listed in the present study, they may think more carefully about design and analysis issues from the beginning of the study, thus raising the overall quality of research (23, 34). Epidemiologists and biostatisticians may be more prone to report these features because of their training (44), which may partially explain why the articles in epidemiology journals in this study reported the most checklist criteria. Journal policy toward reporting observational longitudinal research can clearly contribute. A review of authors' guidelines for the six journals used in this study showed a rather low level of required detail specific to nonrandomized designs. The reporting of methodological detail about aspects that threaten internal validity are the domains of editors (and journal policy) and authors. Higher journal quality indicators, such as impact factors, have been linked to better overall reporting in randomized and nonrandomized studies (45); however, we failed to show a clear trend in this study.

We developed a flow diagram that summarizes sample selection, participant recruitment, eligibility criteria, consent and reasons for nonconsent, timing of follow-ups, and attrition at each stage. The choice of criteria to include was based on the desire to capture the key aspects that allow editors and readers to rapidly judge threats to the internal and external validity of the study, balanced with the need to keep the diagram relatively simple. Detail about the analysis was not included to avoid complicating the diagram. As expressed by Rennie, commenting on the benefits of CONSORT, "[when using a] … checklist and flow diagram, it takes a fraction of the time to get the essential information necessary to assess the quality of a trial" (46, p. 2006).

We recommend that editors move to require authors to use a structured approach to presenting the architecture of observational longitudinal research to communicate essential details about the study design. Doing so may force researchers to organize their thinking during an early stage of their research. The combination of a checklist such as ours, a flow diagram, and, ideally, a structured abstract (47) offers a starting point for consideration.

ACKNOWLEDGMENTS

Drs. L. Tooth and R. Ware were supported by a National Health and Medical Research Council of Australia Capacity Building Grant (252834).

Assistance is acknowledged from Drs. A. Barnett, Z. Clavarino, J. Najman, A. Lopez, P. Schluter, G. Williams, J. Van Der Pols, A. Mamun, and R. Alati from the Longitudinal Studies Unit at the School of Population Health, University of Queensland (URL: http://hisdu.sph.uq.edu.au/lsu/); and from Dr. A. Green from the Queensland Institute of Medical Research.

REFERENCES

1. Reiffenstein R, Schiltroth A, Todd D. Current standards in reported drug trials. Can Med Assoc J 1968;99:1134–5.
2. Freiman J, Chalmers T, Smith H Jr, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690–4.
3. DerSimonian R, Charette J, McPeek B, et al. Reporting on methods in clinical trials. N Engl J Med 1982;306:1332–7.
4. Schulz K, Chalmers I, Hayes R, et al. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273:408–12.
5. Moher D, Schulz K, Altman D, et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285: 1987–91.
6. CONSORT. (http://www.consort-statement.org).
7. Devereaux P, Manns B, Ghali W, et al. The reporting of methodological factors in randomised controlled trials and the association with a journal policy to promote adherence to the consolidated standards of reporting trials checklist. Control Clin Trials 2002;23:380–8.
8. Moher D, Jones A, LePage L, et al. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before and after evaluation. JAMA 2001;285:1992–5.
9. Egger M, Juni P, Bartlett C, et al. Value of flow diagrams in reports of randomised controlled trials. JAMA 2001;285:1996–9.
10. Des Jarlais D, Lyles C, Crepaz N, et al. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004;94:361–6.
11. Moher D, Cook D, Eastwood S, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Lancet 1999;354:1896–900.
12. Stroup D, Berlin J, Morton S, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA 2000;283:2008–12.
13. Bossuyt P, Reitsma J, Bruns D, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. AJR Am J Roentgenol 2003;181:51–5.
14. Ioannidis J, Haidich A, Lau J. Any casualties in the clash of randomised and observational evidence? BMJ 2001;322:879–80.
15. National Health and Medical Research Council of Australia. How to review the evidence: systematic identification and the review of the scientific literature. Canberra, Australia: Biotext, 2000.
16. Grimes D, Schulz K. Cohort studies: marching towards outcomes. Lancet 2002;359:341–5.
17. Vandenbroucke J. Observational research and evidence-based medicine: what should we teach young physicians? J Clin Epidemiol 1998;51:467–72.
18. McKee M, Britton A, Black N, et al. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ 1999;319:312–15.
19. Glasziou P, Vandenbroucke J, Chalmers I. Assessing the quality of research. BMJ 2004;328:39–41.
20. Rychetnik L, Frommer M, Hawe P, et al. Criteria for evaluating evidence on public health interventions. J Epidemiol Community Health 2002;56:119–27.
21. Sackett D, Wennberg J. Choosing the best research design for each question. (Editorial). BMJ 1997;317:1636.
22. Epidemiology Work Group of the Interagency Regulatory Liaison Group. Guidelines for documentation of epidemiologic studies. Am J Epidemiol 1981;114:609–13.
23. Wolfe F. Critical issues in longitudinal and observational studies: purpose, short versus long-term, selection of study instruments, methods, outcomes and biases. J Rheumatol 1999;26: 469–72.
24. Hartz A, Marsh J. Methodologic issues in observational studies. Clin Orthop 2003;413:33–42.
25. Cho M, Bero L. Instruments for assessing the quality of drug studies published in the medical literature. JAMA 1994;272: 101–4.
26. Boulware L, Daumit G, Frick K, et al. Quality of clinical reports on behavioural interventions for hypertension. Prev Med 2002;34:463–75.
27. Thakur A, Wang E, Chiu T, et al. Methodology standards associated with quality reporting in clinical standards in pediatric surgery journals. J Pediatr Surg 2001;36:1160–4.
28. Rushton L. Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. Occup Environ Med 2000;57:1–9.
29. Kleinbaum D, Morgenstern H, Kupper L. Selection bias in epidemiologic studies. Am J Epidemiol 1981;113:452–63.
30. Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol 1977;106:184–7.
31. Zapf D, Dorman C, Frese M. Longitudinal studies in organizational stress research: a review of the literature with reference to methodological issues. J Occup Health Psychol 1996;1:145–69.
32. Wolfe F, Lassere M, van der Heijde D, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26: 484–9.
33. Zaccai J. How to assess epidemiological studies. Postgrad Med J 2004;80:140–7.
34. Grimes D, Schulz K. Bias and causal associations in observational research. Lancet 2002;359:248–52.
35. Rothman KJ, Greenland S, eds. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven, 1998.
36. Hennekens C, Buring J. Epidemiology in medicine. Boston, MA: Little, Brown & Company, 1987.
37. Last JM, ed. A dictionary of epidemiology. 4th ed. New York, NY: Oxford University Press, 2001.
38. Dunn K, Jordan K, Lacey R, et al. Patterns of consent in epidemiologic research: evidence from over 25,000 responders. Am J Epidemiol 2004;159:1087–94.
39. Desmond D, Bagiella E, Moroney J, et al. The effect of patient attrition on estimates of the frequency of dementia following stroke. Arch Neurol 1998;55:390–4.
40. Reijneveld S, Stronks K. The impact of response bias on estimates of health care utilization in a metropolitan areas: the use of administrative data. Int J Epidemiol 1999;28:1134–40.
41. Twisk JW. Applied longitudinal data analysis for epidemiol-

ogy: a practical guide. Cambridge, United Kingdom: Cambridge University Press, 2003.
42. Hankinson SE, Colditz GA, Hunter DJ, et al. A prospective study of reproductive factors and risk of epithelial ovarian cancer. Cancer 1995;76:284–90.
43. Dijkers M, Kropp G, Esper R, et al. Reporting on reliability and validity of outcome measures in medical rehabilitation research. Disabil Rehabil 2002;16:819–27.
44. Delgado-Rodriguez M, Ruiz-Canela M, De Irala-Estevez J, et al. Participation of epidemiologists and/or biostatisticians and methodological quality of published controlled clinical trials. J Epidemiol Community Health 2001;55:569–72.
45. Lee K, Schotland M, Bacchetti P, et al. Association of journal quality indicators with methodological quality of clinical research articles. JAMA 2002;287:2805–8.
46. Rennie D. CONSORT revised—improving the reporting of randomized trials. (Editorial). JAMA 2001;285:2006–7.
47. Haynes R, Mulrow C, Huth E, et al. More informative abstracts revisited: a progress report. Ann Intern Med 1998;113:69–76.