International Journal of Epidemiology
© International Epidemiological Association 1987
Vol. 16, No. 2
Printed in Great Britain

Recall Bias: A Proposal for Assessment and Control

KAREN RAPHAEL

Probably every major text on epidemiology offers at least some discussion of recall bias in retrospective research. A potential for recall bias exists whenever historical self-report information is elicited from respondents. Thus, the potential for its occurrence is greatest in case-control studies or cross-sectional studies which include retrospective components. Recall bias is said to occur when accuracy of recall regarding prior exposures is different for cases versus controls.

The current paper will describe the process and consequences of recall bias. Most critically, it will suggest methods for assessment and adjustment of its effects.

WHAT IS RECALL BIAS?

Generally, recall bias has been described in terms of 'embroidery' of personal history by those respondents who are cases. For example, Mausner and Kramer[1] comment that 'people may be more likely to search for explanations for the disease in the cases and, therefore, may assign more significance to past events' (p. 165). A classic example is reported by Brown and Harris[2] who comment on a 1958 study by Stott.[3] Stott, finding that mothers of 'mongol' children reported more shocks during pregnancy than mothers of normal children, concluded that socioemotional factors had an aetiological role in mongolism. Later research, of course, indicated that chromosomal abnormalities were the cause, and that ordinary events were probably redefined as traumas by mothers of cases in an effort to explain what happened.

Perhaps less recognized as part of the recall bias process is that those without the disease under investigation (ie controls) may be less likely to recall a true exposure. For example, mothers who have given birth to healthy infants may have less motivation to recall earlier drug use during pregnancy than those mothers whose infants had birth defects.

Psychiatric Epidemiology Training Program, Columbia University, School of Public Health, Tower 3-20E, 100 Haven Avenue, New York, NY 10032, USA.

Whether the source of the bias is underreporting of true exposures in controls or overreporting of true exposures in cases, the net effect of the bias is to exaggerate the magnitude of the difference between cases and controls in reported rates of exposure to risk factors under investigation. Consequently, recall bias leads to an inflation of the odds ratio. It leads to the likelihood that significant research findings based upon retrospective data can be interpreted in terms of a methodological artefact rather than substantive theory.

It is important to note that recall bias is *not* equivalent to memory failure itself. If memory failure regarding prior events is equal in case and control groups, recall bias will not occur. Rather, memory failure itself will lead to measurement error which, in turn, will usually lead to a loss of statistical power. Loss of power will bias hypothesis tests toward the null. In contrast, inequality of memory failure in cases versus controls leads to recall bias which, in turn, will bias hypothesis tests away from the null.

The confusion between simple memory failure and differential memory failure is common in the literature. For example, a 1967 study by Klemetti and Saxen[4] is often cited as evidence that recall bias exists in retrospective studies of maternity outcome. However, as noted by Lippman and Mackenzie,[5] Klemetti and Saxen found that, while mothers' retrospective reports of drug use during pregnancy poorly corresponded with their prospective reports, the degree of memory loss was not moderated by the outcome of the pregnancy or condition of the child. Thus, there was no support for anamnestic inequality between cases and controls. A similar confusion between simple recall difficulty and recall bias occurred in a recent report.[6] When comparing data on accuracy of two methods for collection of data on perimenstrual stress, the authors note the unreliability of retrospective reports and state that their results are consistent with recall bias hypotheses. However, they did not note differential unreliability among cases and controls and, therefore, did not support recall bias hypotheses.

Downloaded from https://academic.oup.com/ije/article/16/2/167/668681 by guest on 23 April 2024

Despite the attention given to recall bias in most epidemiological texts, the research literature focusing specifically on recall bias is extremely limited. A computerized literature search on the topic scanned four data bases (ie *Psychological Abstracts* post-1967, *Biological Abstracts* post-1977, *Index Medicus* post-1966 and *Science Citation Index* post-1977) and found only 23 articles mentioning recall bias in their title or abstract. Five of these 23 articles were general methodological papers on the case-control study. Seven were critical reviews of epidemiological research for a specific disease. Seven were studies in experimental psychology of memory and cognition, and only some of these had application to recall bias in epidemiological research. Finally, only four were specific original research papers. Three of these four papers were in the field of maternity outcome. An additional manual search of the research literature suggests that case-control reports of maternity outcomes are more likely than other areas of research to discuss recall bias. Recall bias appears to have been given somewhat less attention in other areas of epidemiological research.

No claim can be made that a computerized literature search represents an all-inclusive search of articles discussing recall bias. While many more original articles probably mention issues of recall bias in their discussion sections, the fact that so few studies mention recall bias in their abstract suggests that the issue is ultimately accorded limited *prominence* in the conclusions reached from the studies.

Nevertheless, the same texts which recognize the possibility of recall bias in case-control studies usually note that the case-control design is the most appropriate methodology for study of rare or chronic diseases with long incubation or latency periods. Case-control designs are also most appropriate for generating hypotheses in newer areas of research and when working under tight budgetary restrictions. Thus, although problems posed by recall bias may be substantial, the necessity for case-control designs suggests that these problems will be with us for some time to come.

While firm evidence for the presence of recall bias is inconclusive,[5] multiple studies have demonstrated its existence in some areas of psychiatric research.[7] A conservative approach in all areas of research would dictate that a study be viewed as 'guilty until proven innocent'. It is not the task of one's critics to prove that recall bias exists; rather, I contend that it is the researcher's task to either present a strong case against the existence of this threat to the study's validity and/or try to statistically control for recall bias in one's analysis.

METHODS OF ASSESSING AND ADJUSTING FOR RECALL BIAS

Given the possibility that recall bias may be operating, it becomes imperative to try to assess its effect in any case-control or cross-sectional study.

Recall bias can be ruled out when actual exposure status can be verified through unbiased records. Since no unbiased records are available in most studies, this approach has only limited application. In fact, if unbiased records are available, it would often seem less expensive and more accurate to avoid the questioning of respondents at all regarding the verifiable exposures.

Recall bias is less likely to be invoked as a plausible explanation for research findings when other exposures have equal 'intuitive plausibility' as risk factors for the disease under investigation, in comparison to the significant exposure, yet the other exposures are not significantly related to disease status. For example, in a case-control study on maternal use of Bendectin and infant malformation,[8] a significant odds ratio was found for Bendectin use. No significant effects were found for other drugs used during pregnancy. The authors convincingly argue against the possibility of recall bias, because biased recall of exposure to a potential teratogen should be equal across the different agents.

Another possible method for assessing whether recall bias is operating at all is to directly ask respondents to identify the exposures which they believe are relevant risk factors for the disease. To reduce sensitization, the questioning should occur following the completion of the rest of the interview. If those risk factors which appear to be significantly related to case status are not believed by respondents to be risk factors for the disease at all, one possesses fairly good evidence against the existence of recall bias. However, what strategy can one apply in a study where the investigated exposures *are* plausible risk factors to the respondent?

The strategy I propose is adapted from the logic of the validity scales of the Minnesota Multiphasic Personality Inventory (MMPI),[9] the most popular personality inventory to date.[10] The authors of the MMPI included in their multi-scale instrument a validity scale to correct or adjust some of the other scales, based upon a measure of each respondent's test-taking attitude or response set.

Similarly, a validity scale can be constructed to adjust or correct for differential recall patterns among respondents. Construction of the validity scale will involve the researcher's identification of a number of previously evaluated exposures which have been *ruled*

Downloaded from https://academic.oup.com/ije/article/16/2/167/668681 by guest on 23 April 2024

out as risk factors for the disease under study. Each item on the validity scale should question the respondent regarding his or her past exposure to a specific 'fake' risk factor. When case respondents positively endorse an excessively large number of validity scale items in comparison to control respondents, it is likely that the endorsement is due to overreporting recall bias rather than actual higher rates of exposure.

The exposures on the validity scale should be of approximately equal plausibility when compared to the exposures which are the putative risk factors of the research project. If exposures are not of equal plausibility, the validity scale will not appropriately measure 'search for cause' cognitive processes. It would be important to verify the equal plausibility of 'fake' validity scale exposures versus 'true' exposures of research interest by asking a small additional group of respondents (who are similar in composition to the larger case and control groups) to rate all exposures according to the likelihood that each exposure is a risk factor for the disease under investigation.

If previous research on the outcome of current interest is sufficient in scope, sophistication, and diversity, it would also be helpful to identify specific exposures which appeared to be risk factors in initial retrospective studies but which were ruled out as risk factors in later prospective research. These exposures would probably be the exposures most sensitive to influence through recall bias.

By comparing total validity scale scores for cases versus controls, it will first be possible to assess whether and to what extent recall bias is occurring. Most importantly, since the validity scale score is a function of the extent of each respondent's recall bias, it may be entered into the final analysis as a statistical control for recall bias.

The inclusion of irrelevant exposures in the research protocol will, in itself, help to reduce recall bias by reducing the likelihood of hypothesis-guessing by either the interviewer or the respondent. Therefore, it will aid efforts to keep both interviewer and respondent blind to research hypotheses.

When studying certain disease outcomes, there may be little or no previous research which allows us to identify plausible exposures which have been ruled out as risk factors for the disease. Another type of validity scale could then be constructed, based upon exposure information for which there is independent verification on an individual respondent basis. Usually, medical records may be used. These exposures may or may not be exposures which the study intends to investigate substantively. The discrepancy between respondent reports and validated record reports can be used as another estimate of the extent to which respondent reports are biased by recall. Again, the discrepancy scores, summed across a number of items, can be used in the final statistical analysis as a control for recall bias effects. Odds ratios resulting from the combined analysis of risk factors and the validity scale can be statistically adjusted for recall bias effects.

The latter method of validity scale construction need not be confined only to studies where a set of plausible but 'fake' risk factors cannot be identified. If record verification is possible, it is recommended that the latter scale construction method be used in combination with a method based upon reports of exposure to plausible but false risk factors.

LIMITATIONS OF THE VALIDITY SCALE APPROACH

The validity scale strategy is not offered as a final panacea for the problem of recall bias. Its success and utility will be dependent upon the ability to locate equally plausible but false risk factors or to verify selected exposures through independent records. In many studies, it may not be possible to meet either of these conditions.

On a practical level, inclusion of 'irrelevant' exposure histories in a research protocol will add length to the interview. The added interview length will undoubtedly add cost to the study and may reduce participation rates. From a cost-benefit perspective, there must be a trade-off between internal validity benefits versus budgetary and generalizability costs.

Finally, there is remarkably little research which systematically evaluates either whether or when recall bias is likely to be a problem. Future utilization and evaluation of the proposed validity scale approach may help to provide critical data on these issues.

REFERENCES

[1] Mausner J S, Kramer S. *Mausner & Bahn—Epidemiology: An Introductory Text.* Philadelphia, W B Saunders, 1985.

[2] Brown G W, Harris T. *Social Origins of Depression: A Study of Psychiatric Disorder in Women.* New York, Free Press, 1978.

[3] Stott D H. Some psychosomatic aspects of causality in reproduction. *J Psychosom Res* 1958; **3**: 42–55.

[4] Klemetti A, Saxen L. Prospective versus retrospective approach in the search for environmental causes of malformations. *Am J Public Health* 1967; **57**: 2071–5.

[5] Lippman A, MacKenzie S G. What is recall bias and does it exist? In M Marois (Ed) *Progress in Clinical and Biological Research, Vol. 163. Prevention of Physical and Mental Congenital Defects: Part C: Basic and Medical Science, Education, and Future Strategies.* New York, Alan R Liss, 1985.

[6] Woods N F, Most A, Dery G K. Estimating perimenstrual distress: A comparison of two methods. *Res Nurs Health* 1982; **5**: 81–91.

[7] Wolkind S, Coleman E Z. Adult psychiatric disorder and childhood

Downloaded from https://academic.oup.com/ije/article/16/2/167/668681 by guest on 23 April 2024

experiences: The validity of retrospective data. *Br J Psychiat* 1983; **143:** 188–91.

[8] Eskenzi B, Bracken M B. Bendectin (Debendox) as a risk factor for pyloric stenosis. *Am J Obstet Gynecol* 1982; **144:** 919–24.

[9] Dahlstrom W G, Welsh G S. *An MMPI Handbook: A Guide to Use in Clinical Practice and Research.* Minneapolis, University of Minnesota Press, 1960.

[10] Anastassi A. *Psychological Testing.* New York, MacMillan, 1976.

Downloaded from https://academic.oup.com/ije/article/16/2/167/668681 by guest on 23 April 2024