Reliability of Morbidity Data in Family Practice

John E. Anderson, MD
Kingston, Ontario

Because of its relative youth, family practice research has not yet developed a tradition of proven research techniques. New techniques, even those already proven effective in other disciplines, must be evaluated in the family practice setting if the results that they generate are to have any credibility. The collection of morbidity data has become a major activity in family practice research, but this has occurred without sufficient examination of its reliability. Several problems, both potential and real, exist requiring more detailed scrutiny, discussion, and possibly action. These problems of recording, diagnosis, coding, and population, and their ramifications, are explored with the aim of stimulating such action and encouraging a rigorous approach to the collection, publication, and interpretation of morbidity statistics.

Just as family medicine is a relatively new academic discipline, so is family practice research a relatively new activity. Both the discipline and its research arm have been the subject of discussions about the knowledge base that they should teach and study. Geyman has described the scope of potential research in and into family medicine.[1] Although this huge, exciting panorama creates considerable enthusiasm, it must be approached with caution.

As a new field, family practice research has not developed a stock of tried and proven methods. New tools are necessarily being developed while others are being adopted and/or adapted from other disciplines, notably epidemiology and the behavioral sciences. Caution is required to ensure that these new or borrowed methods function accurately in the family medicine setting and that they are applied appropriately. The reliability and precision of each method must be evaluated if the results of its employment are to be credible.

For example, difficulties with the denominator problem and with morbidity classification have been recognized already and are under review,[2-11] but another activity appears to have achieved major prominence in family practice without sufficient consideration of its validity and reliability. This is the collection of diagnostic information for the compilation of morbidity statistics. Descriptive reports of disease distribution are seen as useful in examining differences in morbidity patterns between areas or jurisdictions, but their use—and the interpretations that flow from them—are based on the assumption that they are reliable. This assumption deserves critical attention because con-

From the Department of Family Medicine and Community Health and the Department of Epidemiology, Queen's University, Kingston, Ontario. Requests for reprints should be addressed to Dr. John E. Anderson, Department of Family Medicine, 220 Bagot St, PO Bag 8888, Kingston, Ont., Canada K7L 5E9. This project was undertaken while the author was on leave as a Faculty Scholar of the Josiah Macy, Jr. Foundation.

0094-3509/80/040677-07$01.75

tinuing unquestioned reliance on potentially faulty data can harm the whole of family practice research and not only the individual studies involved. Only when deficiencies have been sought, recognized, and assessed can the reliability of morbidity statistics be determined. This paper will review several real and/or possible deficiencies as an initial step in this process.

Mortality Statistics

Morbidity statistics are really an outgrowth of mortality statistics and, as such, they should be examined both for the known weaknesses of their progenitor and for any inherent flaws. There are certainly proven deficiencies in mortality data. There are problems in the registration process itself, in the application of rules for selecting the cause of death from a list of diagnoses, in diagnostic accuracy, and in coding.[12-16]

Mortality statistics are based on a well-circumscribed event, death, and on diseases that are usually definable and, thus, diagnosable with some accuracy. In contrast, morbidity statistics are based on what is usually a continuing event of variable length and with blurred onset and termination—the illness "episode." These latter events are more often poorly defined and less amenable to accurate diagnosis. Given that there is a certain level of imprecision in the more precise area of mortality, what is the level of the problem in the realm of morbidity?

Problem Areas in Morbidity Statistics

Any problems with morbidity statistics, whether real or potential, are significant only to the extent that they affect the purposes for which the data are to be used. Obviously, the collection of information for morbidity registers cannot be ignored, but the major emphasis in this discussion will be on publication of morbidity rates.

The major purpose of such publications is for comparison, for demonstrating similarities or differences between groups or areas. Such a comparison demands that the sources have a certain basic level of similarity. In a case control study, the controls must match the cases as closely as possible, except for the study variable. So too with comparisons of morbidity data. If there are too many confounding variables (differences between the data sources), then it is virtually impossible to determine the role of the study factor (eg, geographic location) in causing differences in rates.

Problems with morbidity statistics fall into four groups: problems of recording, diagnosis, coding, and population. Many of the problem areas are potential, rather than proven. Such is the current state of the art that the significance, indeed the reality, of some of these problems cannot be assessed.

Recording

1. The *purpose of the recording* may have some effect on the reliability of the results. For example, bias may be introduced if a prime purpose of a system is to facilitate billing procedures. Since the purpose of reporting is to justify payment of a fee, that justification will be paramount and accuracy may become a secondary consideration. Physicians may substitute more medically acceptable diagnoses for some social situations, such as housing problems, to justify payment from a "medical" insurance scheme. They may also make substitutions for other diagnoses, such as venereal diseases or therapeutic abortion, to keep such confidential information out of the hands of third parties. Obviously these substitutions lead to underreporting of the problems in question and to overreporting of the substituted labels. By the same token, physicians may record only one of multiple illnesses dealt with, since one diagnosis is sufficient for payment. These possible difficulties dictate special caution when examining data from the files of medical insurance plans.

2. The *frequency of recording* undoubtedly has an effect on the quality of the final data. Systems in which physicians report the same problem every time that they see it are more prone to inconsistent reporting and therefore to error. The likelihood of having the same illness reported under more than one diagnostic heading will result in an underemphasis of the correct diagnosis and an overemphasis on the others. The reporting of every encounter with patients having hypertension

was one factor that led to an error of more than 50 percent in the apparent prevalence of hypertension in one data set.[11] On the other hand, physicians who only report an illness once during its course run the risk of forgetting to report some episodes.

3. The *recorders themselves* cannot help but have an effect on the resultant picture of illness distribution. Thus, one should not attempt to compare data from centers that include reports from nurses and social workers with data from centers that specifically exclude such sources from their reporting. The data bases will obviously be different.

Differences between individual physicians can cause apparent incongruities in resultant morbidity distributions. The dedication of individual reporters to their task, and the volume of the rest of their workload can affect the accuracy and completeness of their reporting. In one family practice residency program, it has been demonstrated that residents actually reported, on average, one problem less than was actually dealt with.[17] Recording losses may well be higher in less motivated settings. Physicians having special clinical interests report higher frequencies of morbidity within those spheres of interest,[18] whether from the provision of consultative services, heightened sensitivity, or diagnostic prejudice.

4. The *geographic location* in which the physician practices has a bearing on reported morbidity distributions. Urban/rural differences are a case in point. Three Canadian studies have shown that urban physicians report higher rates of emotional illness than their rural colleagues.[18,19,20] Other differences were less consistent. This potential for regional variation was recognized in recruiting recorders for the National Morbidity Surveys in Great Britain.[21]

While this particular factor is often the study variable in comparing two sets of morbidity data, the two sets must come from similar sources to make the comparisons valid. For example, no conclusions on national differences in Canadian and British illness rates should be based on a comparison of data from inner London with data from rural and remote northern Ontario.

5. The *physical setting* from which the report is generated will alter the type of morbidity reported. Reports from groups that have a high work load in student health services, emergency departments, chronic institutions, and industry will be at variance with data coming from practices that do not have similar involvement. Because of the special nature of morbidity requiring service outside the office, eg, home visits, those systems that report only office encounters will be biasing their results towards the under-reporting of some diseases.

6. The *date of recording* is important. Diagnostic fashions change over time, whether because of new treatments, new information, or new emphasis. This temporal influence may have more effect on reported differences in rates than do actual changes in morbidity.

The two British national morbidity surveys contain examples of this phenomenon. Between surveys, the rate for hay fever doubled; that for the common cold increased by 25 percent; for acute sinusitis, the rate increased by 600 percent, and that for depressive neurosis by a startling factor of 22.[21] Objective analysis led to the conclusion that these changes were due to factors other than a real increase in the level of morbidity.[21]

7. The *number of diagnoses recorded* at an encounter will also have an effect on the final statistics. Those physicians that record only one diagnosis per encounter will show a lower total rate of morbidity in their practices. Since they usually record only the most important condition dealt with on any one encounter, this under-reporting will be selectively biased towards the minor illnesses. Further, Bentsen found that experienced physicians disagreed on the major diagnosis in 15 percent of cases.[17] If only the major diagnosis is being recorded, this level of disagreement would result in important differences in ultimate data sets.

8. *Continuity of care* will encourage consistency of recording for the same problem in the same patient. Lack of such continuity was probably another factor that caused the problem with hypertension rates alluded to earlier.[11]

9. The *act of recording* may affect the accuracy of the report. Morrell has noted that "morbidity studies in some way constrain the doctor to make a diagnosis,"[22] ie, to label a collection of symptoms with a definitive diagnostic tag of questionable veracity.

10. The accuracy of reports is inversely proportional to the *interval between service and recording*. It will, indeed, be the rare physician who does not remember leaving his patient records, in-

surance forms, or morbidity reports just a little too long to be totally positive about all of the details of a patient visit.

11. The *recording system* itself can have much to do with the accuracy and completeness of the reports. Systems that require little additional effort, that are seen as a part of a routine, and that have some perceived benefit to the recorders are likely to contain the more reliable data.

Diagnosis

1. *Diagnostic criteria* are poorly established for many of the more common problems encountered in family practice. The effect of this deficiency on morbidity statistics is clear. How can one really compare the incidence of an illness between two practices if one cannot be certain that the diagnostic label means the same thing in the two groups? This is a particular problem in the field of psychosocial illnesses. There is absolutely no guarantee that the diagnosis of anxiety neurosis means the same thing to different physicians, even if they practice in the same building.

During one 12-month period at this center, the prevalence rate of anxiety neurosis among females aged 15 to 64 years was 95.8/1,000 attending patients (of the same age and sex). The corresponding rate for men was 46.8. The female excess (by a ratio of 2.1:1) is quite in keeping with other studies,[18,23-27] but is the between-sex difference real?

How many of the recording physicians have definitive criteria for the diagnosis of anxiety neurosis? Probably very few, and even among those few, there is no assurance that the criteria are similar. Until such problems are dealt with, it will not be possible to look for the reasons behind the excess of reported psychiatric morbidity among women. Men may be presenting with similar problems, but being diagnosed as "chest wall pain" or "fatigue NYD" (not yet diagnosed).

To take another example, some physicians insist on obtaining a mid-stream urine culture with a colony count in excess of 10^5 before they will diagnose a urinary tract infection. Other physicians are content with a careful microscopic examination of the urine. Others again are less rigorous. What factors are really being compared when frequencies from these practices are studied? Differences in published incidence and prevalence rates may well be the result of physician factors and not of patient or population factors.

2. The *level of diagnosis* is a closely related problem. Where the difference has no real clinical significance, many physicians are satisfied to report manifestational diagnoses, eg, "tension headache" or "anxiety," as opposed to etiological diagnoses, eg, "sick child" or "marital problem." This aspect of clinical medicine is highly individualistic and there is no way of correcting the biases that it is bound to introduce once it becomes a part of a data reporting system. It can only be avoided by a prior agreement on the level of diagnoses to be reported, and an ongoing monitoring to be sure that the agreement is being lived up to—a cumbersome process. Perhaps pooling of results from several physicians will have some effect on smoothing out the differences, but the larger the number of recorders, the greater the difficulty in standardizing the data collected.[22]

3. The *importance of the diagnosis* may well affect the accuracy of the report. For many physicians, diagnostic accuracy is only important to the extent that it will assist them in helping the patient. Thus, for a self-limited illness of the respiratory tract, different physicians may label the same illness as "influenza," "bronchitis," "tracheitis," "pneumonia," or even "viral illness NYD." Howie has shown the relatively greater significance of signs and symptons, as compared to diagnosis, in the management of some respiratory illnesses.[28] If a physician can decide on the necessary treatment before gathering enough data to establish a firm diagnosis, diagnostic accuracy may suffer, although the patient will not.

4. Following this line of reasoning, *therapeutic decisions* may affect the diagnosis, rather than vice versa. Once again, reference to the British surveys will provide an example. A drop in the rates for menopausal symptoms paralleled the rise in neurotic depression, suggesting that the diagnosis of neurotic depression may have been used as an alternative diagnosis in the second survey.[21] This hypothesis was substantiated by the age and sex specific incidence rates for neurotic depression. If real, could this substitution have arisen because physicians perceived antidepressant therapy as more beneficial and/or safer than estrogen treatment?

Another possible example comes from data arising out of an unpublished study of the effect of patient gender on tranquilizer prescribing.

Prescribing rates to men and women were reviewed for six psychiatric diagnoses and three psychosocial diagnoses. It was determined that there was a linear relationship between the problem-specific prescribing rates and the prevalence rates of the psychiatric conditions. (The correlation coefficient was 0.98 for men and 0.95 for women.) The two diagnoses that appeared to have the greatest effect in causing this relationship (anxiety neurosis and unspecified anxiety) were the two with the least well-defined diagnostic criteria. Although this finding could well have been the result of chance or bias, other explanations are possible as well. One of the foremost of these must be that the physicians first determined the need for tranquilizer therapy and then assigned a diagnostic tag appropriate to the therapeutic decision. Howie has postulated the same phenomenon wherein the diagnosis "will tend to be a justification for treatment, rather than the reason for it."[28]

5. The *definition of an episode* is confusing[29] but vital in the analysis of morbidity data. Should several related diagnoses be reported as a single illness, or should each be reported in its own right? A child presents with coryza, pharyngitis, and acute otitis media. Is this a single illness? If so, should it be reported with a single diagnosis? If yes, which one? Later the coryza and the pharyngitis resolve and the acute otitis subsides, but a serious otitis lingers. Is this a new illness or a new episode? Probably not, but how should it be reported?

There is also the elderly patient with hypertension. Are the hypertensive heart disease, the congestive heart failure, and the hypertensive retinopathy different problems or all a part of the same illness? Certainly they present quite different management problems to the physician. Therefore, they should probably be reported as separate diagnoses. Unfortunately, there is no convention for dealing with these problems. If all physicians were to report separate diagnoses, or all physicians report only the "root" diagnosis, it would at least be possible to know what is being dealt with and the data could be interpreted accordingly. Probably the current data contain a mixture of approaches, even from individual physicians.

Some agreement is required to be sure that everyone is reporting the same thing. Since family physicians deal with clinical problems, it would seem reasonable to make them the basis of reporting. For these purposes a clinical problem might be defined as any problem that requires individual investigation, therapy, or follow-up.

6. There is probably some confusion over *which diagnosis to report*. McWhinney has demonstrated the existence of a behavioral as well as a clinical diagnosis.[30] Within this model, patients may present identical symptoms for a variety of reasons; particularly significant here are the limits of tolerance and limits of anxiety.

If, for instance, a patient attends with chest pain, not because it is severe, but because he is worried about it, what is the diagnostic outcome? Some physicians will consider the anxiety as the major problem, concentrate their therapy on relieving the concern, and report the diagnosis as anxiety. Other physicians may recognize and deal with the anxiety, but really regard the "disease" as chest pain and report it as such. A third group of physicians may report both anxiety and chest pain, thus creating two illness episodes.

This difficulty arises because physicians do not differentiate between disease and response to disease (behavior). The result is that a certain number of behavioral diagnoses may be contaminating morbidity data. This may be a basic cause of the differences in psychiatric morbidity rates between practices, a difference that has caused some concern.[31] It may also account, in part, for the differences in reported psychiatric morbidity rates between men and women discussed earlier.

7. The *stage of diagnostic resolution* at which recording occurs is significant. The high presentation rate of undifferentiated problems is a hallmark of primary care. Frequently these problems remain nosologically unresolved at the end of the first visit and are recorded accordingly, eg, "cough NYD." If, as often happens, the problem resolves and there are no further visits, there is no difficulty. The diagnostic report of "cough NYD" is an accurate reflection of the illness episode. However, the process of health care frequently continues to a stage of higher resolution. The illness may persist and on a second visit there may be evidence to justify a diagnosis of "viral pneumonia." Not only are there now two diagnostic reports for the same episode, one is highly inaccurate. While this problem can be overcome with close attention to detail in a manual recording system, its management is far more complex in a computerized system. Probably few, if any,

computerized data systems have yet reached the level of sophistication required to overcome this difficulty.

Coding

1. A *standard system of classification* is essential for coding of information that is to be compared between centers. The advent of the ICHPPC has done much to alleviate this problem in family practice research. Unfortunately, even established systems of classification need to be revised periodically. These periodic revisions must be allowed for when comparing data sets collected and coded at different times.[21]

2. The *recombination of subdivision rubrics* can lead to inaccuracies. These are often developed to meet special local needs, then recombined for comparison of data with other centers. This process, however, requires considerable care. In one instance, faulty recombination of rubrics would have resulted in an 11 percent error in the reported prevalence rate of ischemic heart disease.[10]

3. The problem of *coding methods* has been discussed elsewhere.[11] Peripheral coding systems probably have a higher level of coding accuracy than central systems.[21]

4. *Inter-coder variability* may be a problem, despite the relatively concise nature of the ICHPPC. This factor has never been assessed in any detail. At this center, the coding accuracy varies from 92 to 97 percent among the eight members of the secretarial staff who are doing the coding. Unfortunately, very few reports of morbidity data actually mention any assessment of coding accuracy.

Population

"Rates are the hallmark of epidemiology, for they form the basis of comparisons...."[32] To answer questions about causation, differences in disease frequencies, or success of intervention requires the "setting of two rates side by side and making some sense of comparison."[33] Thus far, this paper has discussed the effect that variability in the numerator can have on the feasibility of such a comparison. But rates, by definition, have a denominator as well, and it too can either help or hinder comparisons.

1. Patient *age and sex* are the strongest determinant of morbidity, yet how frequently does one find published descriptions of disease frequency with no mention of the age and sex distribution of the source population? These frequencies are, in fact, nothing more than crude rates and "crude rates must never be used to compare populations of different structure."[34]

This difficulty can be overcome by the relatively simple mathematical techniques of standardization.[15,16,34] This process could, however, be greatly facilitated if there were agreement about a uniform reference population for use with North American primary care data.

2. *Other population factors* are certainly important, but it would be too arduous to standardize for all of them, except in very special circumstances. As a basic rule, the population should be described. If certain variables, eg, race, education, social class, religion, are large enough (or atypical enough) to bias the results, this should be emphasized.

3. *Other denominators* may be more appropriate for some purpose, such as workload studies. They should receive the same rigorous attention as populations to ensure that the data will be comparable.

Discussion

The prime purpose of this paper is to stimulate concern about the reliability of morbidity statistics. It is hoped that this concern will precipitate dialogue and evaluation leading ultimately to resolution of some problems and proof that others are "non-problems." This list of potential weaknesses may not be complete and new ones may be found.

Some solutions are already being developed. The committee responsible for the ICHPPC is drafting a set of diagnostic criteria for each of the classification's rubrics. The level of acceptance of these criteria remains to be seen. Automated coding of data should reduce inter-coder variation within any one center, but variation between centers will be dependent upon their use of the same program or on a rigid comparison and standardization of different methods. The publication of the "Glossary for Primary Care" has provided a provisional beginning to the standardization of operational terms.[35]

Perhaps too, the same fortuitous circumstance will occur in morbidity statistics that has occurred in mortality statistics; that despite inaccuracies on the individual case level, the pooled data will have an acceptable level of reliability.[15]

When not controlled at the stage of data gathering, these problems introduce bias into the results, a bias that cannot always be corrected by post facto mathematical manipulation of the data. Even if the bias is controllable at the analytical stage, it must be recognized before action can be taken.

The possibility of so many sources of error, variability, and confusion in morbidity data should not be used as an argument to abandon their collection and use. Rather, it should be seen as stressing the need for disciplined activity and scientific interpretation. Descriptions of morbidity frequencies are useful for determining similarities and differences in rates. These, in turn, may be the signposts to areas for fruitful research. However, if the source data are not accurate and comparable, there is a major risk that the signposts will indicate only a maze going nowhere.

Finally, morbidity statistics from family practice should be seen for what they are, a reflection of the physicians' diagnostic opinions about the problems that patients bring to them. They are a picture of only a small portion of illness and disability in the community. Even the portion that they represent may be pictured in a biased fashion because of reliance on the process of diagnostic labeling—a highly individualized and often subjective process in primary care.

References

1. Geyman JP: Research in the family practice residency program. J Fam Pract 5:245, 1977

2. Bass M: Approaches to the denominator problem in primary care research. J Fam Pract 3:193, 1976

3. Kilpatrick SJ: On the distribution of episodes of illness: A research tool in general practice. J R Coll Gen Pract 25:158, 686, 1975

4. Garson JZ: The problem of the population at risk in primary care. Can Fam Physician 22:871, 1976

5. Westbury RC, Tarrant M: Classification of disease in general practice: A comparative study. Can Med Assoc J 101:608, 1969

6. Bentsen BG: Classifying of health problems in primary medical care. J R Coll Gen Pract Occasional Paper 1:1-5, 1976

7. International classification of health problems in primary care. Report of the Classification Committee of the World Organization of National Colleges, Academies, and Academic Associations of General Practitioners/Family Physicians. Chicago, American Hospital Association, 1975

8. Froom J: International classification of health problems for primary care. Med Care 14:450, 1976

9. Working Party Report: International classification of health problems for primary care. J R Coll Gen Pract, Occasional Paper 1:6-10, 1976

10. Anderson JE, Lees REM: Optional hierarchy as a means of increasing the flexibility of a morbidity classification system. J Fam Pract 6:1271, 1978

11. Anderson JE: Centralized morbidity coding: International classification of health problems in primary care. Int J Epidemiol 8:257, 1979

12. The accuracy and comparability of death statistics. WHO Chron 21:11, 1967

13. McKenzie A: Diagnosis of cancer of lung and stomach. Br Med J 2:204, 1956

14. Alderson MR, Meade TW: Accuracy of diagnosis on death certification compared with that in hospital records. Br J Prev Soc Med 21:22, 1967

15. Barker DJP, Rose G: Epidemiology in Medical Practice. London, Churchill-Livingstone, 1976

16. MacMahon B, Pugh TF: Epidemiology: Principles and Methods. Boston, Little, Brown, 1970

17. Bensten BG: The accuracy of recording patient problems in family practice. J Med Care 51:311, 1976

18. Anderson JE, Lees REM: Patient morbidity and some patterns of family practice in southeastern Ontario. Can Med Assoc J 113:123, 1975

19. Greenhill S, Singh HJ: Comparison of the professional functions of rural and urban general practitioners. J Med Educ 40:856, 1965

20. Bartel GG, Waldie AC, Rix DB: Rural and urban family practice in British Columbia: A comparison. Can Fam Physician 16(6):121, 1970

21. Trends in national morbidity: A comparison of two successive national morbidity surveys. J R Coll Gen Pract Occasional Paper 3:1-43, 1976

22. Morrell DC: Now and then. J R Coll Gen Pract 29: 457, 1979

23. Morbidity statistics from general practice: Second National Morbidity Survey 1970-1971. In Office of Population Censuses and Surveys: Studies on Medical and Population Subjects, No. 26. London, Her Majesty's Stationery Office, 1974

24. Marsland DW, Wood M, Mayo F: Content of family practice. Part 1: Rank order of diagnoses by frequency; Part 2: Diagnoses by disease category and age/sex distribution. J Fam Pract 3:37, 1976

25. Dixon AS: Survey of a rural practice: Rainy River, 1975. Can Fam Physician 22:693, 1976

26. National ambulatory medical care survey: 1973 summary: United States, May 1973 - April 1974. In National Center for Health Statistics (Rockville, Md): Vital and Health Statistics, series 13, No. 21. DHEW publication No. (HRA) 76-1772. Government Printing Office, 1975

27. Rowe IL: Prescription of psychotropic drugs by general practitioners: Part 1: General. Med J Aust 1:589, 1973

28. Howie JGR: Diagnosis: The Achilles heel. J R Coll Gen Pract 22:310, 1972

29. Eimerl TS, Laidlaw AJ: A Handbook for Research in General Practice. London, E & S Livingstone, 1969

30. McWhinney IR: Beyond diagnosis: An approach to the integration of behavioral sciences and clinical medicine. N Engl J Med 287:383, 1972

31. Warrington AM, Ponesse DJ, Hunter ME, et al: What do family physicians see in practice? Can Med Assoc J 117:354, 1977

32. Rose G, Barker DJP: Epidemiology for the uninitiated: Rates. Br Med J 2:941, 1978

33. Rose G, Barker DJP: Epidemiology for the uninitiated: Comparing rates. Br Med J 2:1282, 1978

34. Hill AB: A Short Textbook of Medical Statistics. London, Hodder and Stoughton, 1977

35. Report of the North American Primary Care Research Group (NAPCRG) Committee on Standard Terminology: A glossary for primary care. J Fam Pract 5:633, 1977