

# Classifying and Coding Morbidity in General Practice: Validity and Reliability in an International Trial

Charles Bridges-Webb, MBBS, MD, FRACGP  
Croydon, New South Wales, Australia

*An international field trial of the draft defined version of the International Classification of Health Problems in Primary Care (ICHPPC) was undertaken in nine countries. Fifty-two physicians coded morbidity in 76 standard written clinical vignettes. The validity and reliability of morbidity coding using the defined classification and the extent to which these might be improved by using definitions have been estimated. There was an average of 1.7 problems per encounter, for a total of 807 different problems recorded. Sixty-two per cent of the problems were coded correctly. The majority of incorrectly recorded problems were due to varied minority views of the contents of the encounters and to mistakes in recording code numbers. The results of morbidity surveys are useful only if the methods of gathering, recording, coding, collating, and analyzing the data are explicit and well documented. This study illustrates the difficulties of recording morbidity in general practice encounters. Better use of definitions and the addition of an index could improve coding accuracy.*

The *International Classification of Health Problems in Primary Care (ICHPPC-2)*<sup>1</sup> was developed by the Classification Committee of the World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians (WONCA). It is compatible with the *International Classification of Diseases (ICD-9)*<sup>2</sup> and contains 362 rubrics in 18 sections, selected because of their frequency, importance, and to a lesser extent, their specificity. The classification is not hierarchical, but the principle of optional hierarchy is accepted.

Using an internationally agreed upon list of rubrics for classifying the problems met in primary care does not in itself ensure the highest possible level of statistical comparability in morbidity surveys.

There are two steps in recording morbidity: naming the problem (formulating the diagnosis), and coding it to a rubric of the classification. The first step is a nor-

mal part of the medical encounter; the second is a requirement of morbidity recording. In general practice naming the problem may involve formulating a specific diagnosis such as measles. It is often possible to name a problem only in a very nonspecific way, in terms of the major presenting symptom or by using a very general diagnostic term, such as viral infection. It is a matter of professional judgment as to whether a more or less specific diagnosis is formulated for a particular problem. This variability is very difficult to control and is a major unresolved source of difficulty in interpreting the results of morbidity surveys.

Having named the problem, the second step in recording morbidity is to assign the named problem to a rubric in the classification. Such assignation is relatively easy if the named problem is exactly equivalent to one rubric in the classification. When the named problem is not exactly equivalent to one rubric, a decision must be made as to which of two or more rubrics provides the best fit. Definitions of rubrics are intended to improve the consistency with which the rubric is chosen and so reduce variability in coding.

Definitions for 282 of the 362 rubrics in ICHPPC-2 were developed. The definitions are stipulative, using the minimum number of discriminatory criteria to re-

Submitted, revised, July 2, 1985.

From the Department of Community Medicine, University of Sydney, Croydon, New South Wales, Australia. Requests for reprints should be addressed to Professor C. Bridges-Webb, Department of Community Medicine, University of Sydney, 11 Croydon Avenue, Croydon NSW 2132, Australia.

duce the chances of miscoding as much as possible. They distinguish conditions only from those other conditions with which they might commonly be misclassified.

The draft definitions determined by international collaboration were field tested in 1981. The field trial included use of the classification to code 76 standard written clinical vignettes devised for the trial.

The use of standard written clinical vignettes allowed estimation of the validity of the use of defined rubrics in standard clinical situations. It also provided information about the reliability of the process by which general practitioners or family physicians perceive and record the clinical content of patient encounters. Despite extensive and increasing use of comprehensive data systems in general or family practice, these issues have as yet been addressed inadequately.

## METHODS

The vignette trial was designed to measure the reliability of recording problems using the definitions. It was also possible, however, to determine the face validity of recorded problems in the vignettes by comparison with a consensus content of each vignette.

The 76 clinical vignettes were selected from a large number submitted from members of the classification committee in many countries. Each was a brief description of one patient encounter, not a clinical record. The vignettes were deliberately but arbitrarily chosen to be representative of the wide range of encounters that occur in general practice and family medicine, brief or long, focused or diffuse, simple or complicated. The problems were symptomatic or diagnostically specific, medical or psychosocial, common or rare, all in many and varying degrees. The completeness of the information included was variable, as in real life.

Some problems were described so that it was deliberately hard to code to the correct rubric unless the definition criteria were strictly followed. The 106 "official" problems (see below) in the vignettes consisted of 45 that could be coded correctly using the condensed title in ICHPPC-2 only, 26 that needed the full description of the rubric, 30 that needed the definition criteria, and 5 that needed detail included only in ICD-9.

In each country participation was invited from general practitioners and family physicians known to be recording or interested in morbidity information. Those who responded could be presumed to be those most interested and well informed. They were given the defined classification in typescript but without an index; many but not all had the ICHPPC-2 book, which includes a list of condensed titles and an alphabetical index; very few had ICD-9. They were asked to

code all problems that they recognized as having been dealt with at the encounter described in each vignette.

This report is based on reports from 52 physicians in nine countries: Australia (8), Canada (3), Denmark (5), Finland (8), Israel (6), Netherlands (10), South Africa (1), the United Kingdom (2), and the United States of America (9).

No official correct morbidity content for each vignette was determined in advance. Results determined independently by two members of the classification committee included 108 and 109 problems, respectively. On review only 106 of these problems were accepted as correct; that is, the clinical information given in the vignette was compatible with the definition of the relevant rubrics. These were designated the official correct problem codes.

On reviewing the results coded by all 52 physicians, a further 48 problems were accepted as correct because (1) they were compatible with the definitions, and (2) they were coded by at least 10 (19 percent) of the recorders. Most of these extra problems were acceptable alternative codes for problems recorded differently in the official result, but 16 were additional problems. There were thus 122 problems deemed correct in the 76 vignettes, 1.6 per vignette.

## RESULTS

The 52 physicians in the trial recorded an average of 129 problems each in the 76 vignettes, 1.7 problems per vignette. A total of 6,692 different problems were recorded.

The extent to which recorders coded in ways similar to the correct results is shown in Table 1. Seventy of the 122 problems (57 percent) were coded correctly by at least 30 physicians (58 percent). Only seven of the problems were coded correctly by fewer than 10 (19 percent) of the recorders; all of these were in vignettes that had been designed deliberately to be difficult.

There was substantial agreement (more than 70 percent) on the coding of 19 vignettes (25 percent) and 54 problems (44 percent). A majority of the 6,692 problems recorded were coded correctly (4,114, 61 percent); 3,388 (51 percent) were official and 726 (11 percent) were acceptable alternatives.

The 2,578 problems (39 percent) that were incorrectly coded fell into two groups: 949 (14 percent) were incompatible with the defined criteria, and 1,629 (24 percent) were not strictly incorrect but represented only a small minority view of the content of the vignette. In most vignettes there was a wide scatter of codes recorded by a very small number of physicians, many of which probably represent mistakes in recording or transcribing code numbers.

Twenty-four problems (20 percent) were incorrectly coded by more than 10 (19 percent) of the physicians. These problems are important because of the fre-

frequency of miscoding, but they were responsible for only 453 incorrectly coded problems, or 7 percent of all problems recorded and 18 percent of the incorrectly coded problems. These errors were more common in vignettes that had been designed to require definition criteria or the detail of ICD-9 for correct coding (Table 2).

Errors in coding seemed to have arisen from several major sources: jumping to conclusions on the basis of inadequate evidence, difficulty in sorting out multiple complaints, difficulties with the classification, overlooking features of the definitions, difficulty in determining the content of follow-up encounters, and missing diagnostic clues.

Patients with multiple complaints, regardless of whether some of these were related to objective signs, presented great difficulty. The available information in one vignette, for example, suggests an underlying neurological disease, but the relative importance of the patient's various symptoms and the nature and importance of the fall are not clear, which makes it hard to know how to diagnose and code. There was no consensus in the trial results, and 23 different codes were used. Even more difficulty was experienced when subjective complaints did not tally with any abnormal physical findings. These difficulties are compounded further when there are also psychological or social problems, as these problems often coexist with other conditions, and in the case of social problems at least, the classification encourages double coding.

Many difficulties with the classification arose from the original decision to keep ICHPPC-2 compatible with ICD-9. This means that some diseases occurring during pregnancy or in the neonatal period are coded differently from their coding at other times, and some infections and congenital conditions are not coded in the appropriate body systems. Some rubrics seem to be inappropriately situated; for example, sunburn with dermatitis rather than with burns, and dermatitis of the eyelids with eye conditions rather than with dermatitis. "Rash with fever" is listed with infections, whereas "rash unspecified" is included with ill-defined conditions; this proved to be another problem in the coding of one vignette. Some terms were hard to locate because they were not itemized in the classification, such as acute retention of urine. Location problems will be overcome with an alphabetical index.

In some instances physicians overlooked features of definitions designed to avoid the very coding error they made. Sinusitis is often a nonspecific term in everyday practice, and this rubric required a definition to distinguish it from nonspecific other upper respiratory tract infection and allergic rhinitis.

Follow-up encounters are often more difficult to code than first encounters because time or treatment may have altered the clinical evidence, so that criteria met earlier in the course of the disease are no longer

**TABLE 1. EXTENT OF CORRECT CODING BY PHYSICIANS**

Number of Physicians Recording Correct Code for the Problem (n = 52)	No. of Problems No. (%)
40+	36 (30)
30-39	34 (28)
20-29	24 (20)
10-19	21 (17)
Under 10	7 (5)
Total	122 (100)

**TABLE 2. EXTENT OF INCORRECT CODING ACCORDING TO REQUIREMENT FOR CORRECT CODING**

Requirement for Correct Coding	No. of Problems	Problems Incorrectly Coded* No. (%)
Condensed title in ICHPPC	45	5 (11)
Full title in ICHPPC	26	1 (4)
Definition criteria or ICD-9	36	12 (33)
Not classified	16	6 (38)
Total	122	24 (20)

*\*Incorrectly coded by more than 10 (19 percent) of physicians*

present. The introduction to the defined classification recognizes that in reporting survey results, it will be necessary to state the method of application of the criteria. In one vignette, for example, it is appropriate to code the rubric "transient situational disturbance, acute stress reaction" on the basis of the history, which meets the criteria, although at the encounter the patient is "back to normal."

Sometimes diagnostic clues to a more specific rubric were missed, and a less specific rubric was coded. In one vignette the reported x-ray evidence allowed a diagnosis of spondylolisthesis to be made, but many recorders coded it as back pain.

## DISCUSSION

Determination of the morbidity content of any physician-patient encounter is generally very imprecise; both the number and the nature of the rubrics that are applied vary considerably. However, this trial shows that in a series of encounters (represented here by vignettes) physicians recorded an average of 1.7 problems per encounter. A similar figure has been found in morbidity surveys<sup>3</sup> and probably reflects the way physicians perceive their work in general practice.

The nature of the morbidity recorded varies much more than the number of problems perceived. Although the average number of problems recorded by the physicians for the 76 vignettes was 129, a total of 807 different patient problems was recorded by the 52 physicians.

Determination of validity depends on having a correct statement of morbidity content for each vignette. Since morbidity content is such a subjective estimation, the correct content has been derived from results determined by two experienced morbidity recorders backed up by committee review and supplemented by acceptable codes determined by consensus of all trial participants.

As preventive and social conditions are part of the problem content of many medical encounters, and the decision to include them as part of the problem content of the encounter is very subjective, it is not surprising that they account for one half of the consensus of additional acceptable codes. Similarly, the overrepresentation of symptoms and ill-defined conditions in the alternative results arises because nonspecific codes are often appropriate to describe morbidity that could also be coded to a more specific rubric.

The difficulties of recording morbidity in general practice encounters are well illustrated in this trial. The reliability of coding is probably better than could be expected. Two experienced committee members achieved consensus for 72 percent of the vignettes and 81 percent of the problems, and among all participants 60 percent of all problems were correctly coded by more than 60 percent of recorders.

The majority of incorrectly recorded problems were due to varied minority views of the contents of encounters and to mistakes in recording or transcribing code numbers; none of these errors will be influenced by definitions of rubrics. Eighteen percent of the incorrectly coded problems (from 25 percent of the vignettes) could be influenced by better use of definitions. Such improvement would be worthwhile if it occurred in morbidity surveys. However, there is no

evidence of how often problems that are difficult to code arise in everyday practice.

The field trial was designed to test the draft definitions. The absence of an index in the typescript classification provided to participants was a considerable handicap. Many of the definitions were modified as a result of the trial, and the published version with a good index may well be even more effective than this trial shows.<sup>1</sup>

Use of a classification, however good and well-defined the rubrics are, can result in useful morbidity data only if the methods of gathering, recording, coding, collating, and analyzing the data are explicit and well documented. It is important that users of the classification be made aware of the need to address these issues and use detailed reference manuals in conducting surveys.

#### **Acknowledgments**

The work was largely undertaken during tenure of a Visiting Professorship in the Department of Family Medicine, State University of New York at Stony Brook, supported by a Special Studies Leave grant from the University of Sydney and grants from the Rockefeller Foundation and the Utah Foundation. All members of the WONCA Classification Committee provided cooperation and assistance in preparing this paper. The 52 practicing physicians who coded the vignettes for the study deserve special thanks.

#### **References**

1. ICHPPC-2 Defined: International Classification of Health Problems in Primary Care, ed 3. Classification Committee of the World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians in collaboration with the World Health Organisation. Oxford: Oxford University Press, 1983
2. Ninth Revision, International Classification of Diseases. In Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death. Geneva, World Health Organization, 1977
3. Royal Australia College of General Practitioners, Bridges-Webb C: The Australian general practice morbidity and prescribing survey 1969-1974. *Med J Aust* 1976; 2(special suppl 1):1-28