Rev Saúde Pública 2008;42(3)

Mariana Miranda Autran Sampaio[I]

Cláudia Medina Coeli[I]

Nair Navarro de Miranda[II]

Eduardo Faerstein[I]

Guilherme Loureiro Werneck[I]

Dora Chor[III]

Cláudia S Lopes[I]

Interobserver reliability of the International Classification of Primary Care

ABSTRACT

OBJECTIVE: The International Classification of Primary Care was developed as an attempt to overcome the limitations of the International Statistical Classification of Diseases and Related Health Problems, 10th revision, when used for primary health care. The aim of the study was to evaluate the interobserver reliability of the International Classification for Primary Care when coding reasons for health-related interruption of daily activities.

METHODS: Data analyzed pertained to 801 subjects from Phase 2 of the Pró-Saúde Study, involving the employees of a Rio de Janeiro university who reported having been prevented from carrying out any of their usual activities (work, study, or leisure) for health-related reasons in the two weeks prior to data collection. Health problems reported in response to an open question were separately coded by two classifiers. Interobserver reliability with respect to number of health problems was calculated by weighted kappa; for the remaining analyses (chapters and full codes), crude kappa coefficients were used.

RESULTS: A total of 1,641 health problems were coded by the first classifier, and 1,629 by the second. Interobserver reliability with respect to the number of health problems coded was substantial (weighted kappa=0.94; 95% CI: 0.93;0.94). Chapter and full codes showed substantial (kappa=0.89; 95% CI: 0.88;0.90) and moderate (0.76; 95% CI: 0.76;0.78) reliability, respectively.

CONCLUSIONS: The results suggest that the International Classification of Primary Care is adequate for the coding of health-related reasons for interruption of daily activities.

DESCRIPTORS: Primary Health Care, classification. Observer Variation. International Classification of Diseases. Questionnaires, utilization. Validation Studies.

INTRODUCTION

Self-reported morbidity surveys are a common tool in health evaluations, especially because they allow access to the morbidity profile of the general population. The population that seeks and obtains health care –usually the object of outpatient and hospital-based studies – may show a profile distinct from that of the general population.[6] One of the issues to be considered when conducting a morbidity survey is the way in which data is collected and coded. Collection can be based both on lists of frequent health problems and symptoms, or on open questions. Studies using the open question without a list of problems or symptoms require a coding strategy. The 10th revision of the International Statistical Classification of Diseases and Health Related Problems (ICD-10)

[I] Instituto de Medicina Social. Universidade Estadual do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

[II] Escola Politécnica de Saúde Joaquim Venâncio. Fundação Oswaldo Cruz (Fiocruz). Rio de Janeiro, RJ, Brasil

[III] Escola Nacional de Saúde Pública. Fiocruz. Rio de Janeiro, RJ, Brasil

Correspondence:
M.M.A. Sampaio
Instituto de Medicina Social
Universidade do Estado do Rio de Janeiro
R. São Francisco Xavier 524, 7º andar
20559-900 Rio de Janeiro, RJ, Brasil
E-mail: m.autran@gmail.com

Received: 11/23/2006
Revised: 10/1/2007
Accepted: 11/8/2007

is the classification system internationally adopted for coding diagnoses and elaborating health statistics. However, this classification is considered inadequate for situations which lack well-defined conditions and where precise diagnosis is impossible, as is the case of health surveys and primary health care.[5,8,a]

The International Classification of Primary Care (ICPC)[15] is an attempt to complement ICD in the context of primary care, and is characterized by the inclusion of patients' complaints and social problems.[3] In order to ensure that comparability between its codes and those of ICD-10, ICPC underwent a revision that became known as ICPC-2. This new version has a biaxial structure: the first axis is divided into chapters that comprise the organic, psychological, and social systems to which the report refer (general and non-specific, blood, digestive, eye, ear, among others); the second axis presents components referring to the type of report (signals and symptoms, procedures, diagnoses, and diseases). The code for a reason for a medical appointment is composed of one letter, which represents the chapter, and two digits, which represent components. Because it was created for primary health care, and because it considers the patient's discourse as it is enunciated, one can expect that ICPC will perform well with open questions contained in health surveys; however, this classification system is recent, and has not been substantially explored in Brazil. In a review of the literature,[b] a single study evaluating the use of this classification in Brazil[7] was identified. Moreover, only one other study used ICPC-2 to code self-reported morbidity in questionnaires, but this study did not provide an evaluation of reliability.[14]

The aim of the present study was to evaluate the interobserver reliability of independent ICPC-2 coding of answers to an open question regarding the reason for interruption of usual activities.

METHODS

The data analyzed in the present study are part of a larger project, the *Pró-Saúde* Study, which has been described in a prior publication (Faerstein et al[2]). That study involves a cohort of technical-administrative employees from a university in Rio de Janeiro state, and was aimed at investigating the association between social determinants and a variety of health-related outcomes. So far, data collection for the *Pró-Saúde* Study was carried out in three phases, in 1999, 2001, and 2006/07. In the present study we evaluated employees that: a) responded to the self-administered questionnaire in Phase 2 (2001); b) provided a valid response to the question on interruption of usual activities due to health reasons; and c) reported having been prevented from carrying out any of their daily activities due to health reasons in the two weeks prior to the interview. Of the total 812 subjects who reported interruptions of their usual activities for health reasons, 11 did not provide the reason for interruption, and were thus excluded from the analysis. Therefore, the population of the present study comprised 801 employees.

The question used to evaluate the interruption of usual activities due to health reasons was the following: "in the last two weeks, which health problem or problems did you have that prevented you from carrying out any of your usual activities (for example, working, studying, leisure activities, or house chores)?" ["*Nas duas últimas semanas, qual foi ou quais foram esses problemas de saúde que você teve ou tem que o(a) impediram de realizar alguma dessas suas atividades habituais (por exemplo, trabalho, estudo, lazer ou tarefas domésticas)*?"]

Coding was carried out by two classifiers: one specialized in disease classification, especially ICD-10, but without prior experience with ICPC-2; and another without any prior classifying experience. Both classifiers went through a training program, consisting of reading of classifications followed by discussion sessions on the logic of the system, prior to beginning the actual classification process. Later, as a test, the two classifiers coded the reported reasons for 29 medical appointments at a primary care unit, as well as 59 questionnaires from the *Pró-Saúde* Study that were not included in the main sample. A meeting was then held in order to arrive to a consensus coding, which involved the participation of a third classifier. This meeting also served to establish directives for coding data pertaining to this study population.[b]

Since any given complaint could be interpreted using more than one code, there was a divergence in the number of reasons codified by each classifier. We therefore carried out a reliability analysis between the two classifiers regarding the number of reasons coded, using the weighted kappa test.[13] In addition, we analyzed the reliability of ICPC-2 according to chapter (for instance, if one classifier coded a reason as P01 and the other as P02, there was agreement in terms of the chapter), according to full code within a given chapter (analyzing separately chapter P, classifiers should agree as to the full code, e.g., P01 and P01), and according to global full code (classifiers should agree as to the full code;

[a] Jamoulle M, Roland M. The WONCA Classification Committee, 1972-1997, 25 years in the service of family practice. 1997 [acesso em 9 nov 2004]. Disponível em: http://www.ulb.ac.be/esp/wicc/history.html

[b] Sampaio MMA. Avaliação da Classificação Internacional de Cuidados Primários na codificação dos motivos de interrupção de atividades habituais no Estudo Pró-Saúde [dissertação de mestrado]. Rio de Janeiro: Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro; 2006.

however, reliability calculation was carried out for the set of all codes, i.e., not taking into consideration the chapter). The analysis of chapters and full codes was subsequently stratified according to sex, schooling, and occupation. Stratification was carried out because subjects, depending on their characteristics, tend to express themselves in different ways, in greater or lesser detail, and using different linguistic peculiarities. Therefore, as self-administered questionnaires do not allow for further clarification, the quality of the report may lead to further difficulty when coding. In all cases, reliability between the two classifiers was estimated using the kappa statistic.[1] 95% Confidence Intervals (95%CI) for the kappa statistic were calculated using the kapci routine,[10] developed for Stata software.

We used the classification proposed by Shrout[11] for interpreting kappa values, as follows: k<0.10 – virtually no reliability; 0.10 to 0.40 – slight reliability; 0.41 to 0.60 – fair reliability; 0.61 to 0.80 – moderate reliability; and 0.81 to 1.0 – substantial reliability. This classification is recently being employed by a number of authors,[12] and represents a further development over the classification proposed by Landis & Koch.[4]

RESULTS

The first classifier codified a total 1,641 reasons, and the second, a total of 1,621 reasons, with both classifiers identifying a median of two reasons per subject. Analysis of the number of reasons codified per subject had a crude agreement of 82.4%, with crude kappa = 0.78 (95% CI: 0.77;0.78) and weighted kappa = 0.94 (95% CI: 0.93;0.94), indicating substantial reliability.

Table 1 shows that reliability estimates for chapter codes were substantial, both globally and within each stratum. When the full code was considered, estimated reliability was moderate to substantial. The small difference related to occupation in the analysis of full codes ceased to exist when only chapter codes were considered. Reliability was also similar for both sexes. However, there were differences relate to schooling, reliability being lower for subjects with less than high-school than for the remainder. Nevertheless, reliability was still considered as moderate among the former.

Table 2 presents the full code agreement level within each chapter of ICPC-2. Reasons included in some of the chapters were infrequent, which hindered the evaluation of reliability for these chapters. Substantial reliability was seen for chapters N (nervous system), R (respiratory tract), and P (psychological). Events tended to be concentrated in a handful of codes, such as headache (178 of 248) in the nervous system chapter; acute upper respiratory tract infection (35 of 183), chronic/acute sinusitis (28 of 183), and flu (52 of 183) in the respiratory tract chapter; and feelings of anxiety/nervousness/tension (71 of 263), acute reaction to stress (31 of 263), and depressive disorders (99 of 263) in the psychological chapter. Chapter K (circulatory apparatus) showed slight reliability. In-depth analysis of discrepancies in chapter K showed that 19 of 41 discrepancies were codified by one classifier as K85 (elevated blood pressure), and by the other as code K86 (uncomplicated hypertension), both with similar meaning. Had both classifiers attributed the same code to such reasons, crude agreement would have risen to 74.4, with a crude kappa of 0.7 (95%CI: 0.63:0.70), considered as moderate.

Table 1. Agreement between classifiers with respect to full code and chapter when classifying reported reasons for medical appointments, according to respondent sex, schooling, and occupation, using ICPC-2. Pró-Saúde Study, Rio de Janeiro, Southeastern Brazil, 2001.

Variable	Chapter		Full code	
	Percent agreement (crude)	kappa (95% CI)	Percent agreement (crude)	kappa (95% CI)
Global	90.6%	0.89 (0.89;0.90)	76.5%	0.76 (0.76;0.78)
Sex				
Male	91.5%	0.90 (0.89;0.91)	78.0%	0.77 (0.74;0.80)
Female	90.3%	0.89 (0.88;0.89)	75.9%	0.75 (0.72;0.79)
Schooling				
Less than high-school	90.0%	0.88 (0.86;0.92)	70.7%	0.70 (0.68;0.72)
High-school	90.8%	0.89 (0.89;0.91)	76.7%	0.76 (0.72;0.78)
College or higher	90.4%	0.89 (0.88;0.89)	78.6%	0.78 (0.75;0.80)
Occupation				
Health professional	89.9%	0.88 (0.87;0.89)	75.2%	0.74 (0.72;0.79)
Other	91.2%	0.90 (0.90;0.92)	77.7%	0.77 (0.76;0.78)

ICPC-2: International Classification of Primary Care, version 2

Table 2. Interobserver agreement with respect to full code when classifying reported reasons for medical appointments according to chapter, using ICPC-2. Pró-Saúde Study, Rio de Janeiro, Southeastern Brazil, 2001.

Chapter	N	Percent agreement (crude)	kappa	95% CI
N (neurological)	248	96.4	0.92	0.86;0.94
R (respiratory)	183	83.6	0.81	0.79;0.85
P (psychological)	263	83.7	0.78	0.75;0.79
H (ears)	18	83.3	0.74	0.53;0.80
U (urological)	21	76.2	0.7	0.62;0.75
D (digestive)	153	71.9	0.69	0.67;0.73
L (musculoskeletal)	407	67.1	0.65	0.61;0.67
X (female genital)	56	69.6	0.65	0.61;0.66
A (general and unspecified)	184	67.4	0.63	0.62;0.65
B (blood)	3	66.7	0.57	0.40;1.00
T (endocrine and metabolic)	18	61.1	0.57	0.44;0.68
F (eyes)	28	60.7	0.56	0.53;0.63
K (circulatory)	86	52.3	0.45	0.41;0.52
Z (social)	7	57.1	0.43	0.00;0.47
W (pregnancy and family planning)	11	36.4	0.34	0.15;0.38
S (skin)	23	30.4	0.29	0.24;0.33
Y (male genital)	0	-	-	-

DISCUSSION

The present study has detected substantial reliability between two classifiers for number of reasons and chapter, and moderate to substantial reliability with respect to full codes. Such high reliability was unexpected, given that one of the classifiers had no prior experience in morbidity coding.

The lower full code reliability of reasons reported by subjects with elementary schooling may indicate that classifiers had difficulty in interpreting the language used by this group. On the other hand, the classification could be made to encompass a wider range of expressions, thus increasing the generalization of its application.

Reliability at the chapter level was similar for health professionals and for other employees, being actually slightly higher among the latter. Health professionals were expected to have greater ability to provide information on their health-related problems, and interobserver reliability was thus expected to be higher for this group. However, this was not the case. A possible explanation for this result is that more detailed reports may actually lead to greater difficulty in classification. For instance, "disc herniation in the cervical spine with canal compromise" ("*hérnias discais na coluna cervical com comprometimento do canal*") was coded by one classifier as "vertebral syndrome," and as "vertebral syndrome with radiating pain" by the other.

The concentration of reasons into a handful of more frequent codes may have contributed to the higher reliability of these chapters. The slight reliability of coding for the circulatory system chapter may be attributed to the existence of different codes for classifying very similar health problems. The distinction between elevated blood pressure and uncomplicated hypertension may be relevant in other contexts, but is not in the case of primary care or health surveys, given that the patient or responder may not have a clear idea of the difference between these two terms, and would be likely to use either one indiscriminately. Notwithstanding, given that this is an important aspect of coding, it will require special emphasis when classifiers are trained.

The only similar study found in the literature was carried out by Letrilliart et al.[9] These authors studied the reliability of the classification of codes attributed by general physicians trained in using ICPC-2, who classified health problems directly from the patient's discourse, and by epidemiologists, who based their classification on medical information extracted from a database. The authors found a weighted kappa coefficient of 0.65 (95% CI: 0.52;0.77) for reliability in the number of reasons codified. Crude agreement at the code level (considering chapter code only) was 69.2% (83 of 120) and crude kappa was 0.84 (95% CI: 0.78;0.91). Even though the comparison of kappa statistics based on different study populations is questionable (given that this measure is influenced by the prevalence of the phenomenon at hand), these values are lower than those estimated in the present study. One may speculate that these discrepancies can be explained, at least partially, by the use of different sources for capturing the data (primary vs. secondary) as well as different professionals (clinicians vs. epidemiologists).

In the study by Van der Heyden et al,[14] ICPC-2 was applied to an open question included in a self-administered questionnaire. These authors reported certain problems with using this classification, including lack of specificity of responses, responses inadequate to the question, cases in which the code to be employed was not clear, and high time demand for coding responses. Lack of specificity was also common in the present study. This may constitute a drawback in applying ICPC-2 to self-administered questionnaires, given that further classification of obscure points in the response is not possible. In the case of primary care, or when questionnaires are completed by the interviewer, there is the possibility of asking further questions in case of unspecific responses. In the present study, there were no cases of responses that were inadequate to the question, nor cases in which the code to be used was not clear.

We found that classification became faster as classifiers became more experienced with the process. However, coding may have taken longer than necessary due to certain problems with ICPC-2, including the lack of expressions in the index such as "allergy," "muscular

distension," "nausea," "flu," "hernia," and "cold," only to list a few. The classification of procedures is also flawed, including terms such as "excision" and "exeresis," but lacking the word "surgery". A classification whose use is not restricted to physicians[7] should carry a wider range of options, so that individuals without prior knowledge of such procedures will also able to work with this system.

In conclusion, ICPC-2 showed good interobserver reliability for coding health reasons for the interruption of usual activities. The fact that the population of the present study comprised the staff of a public university means that it differs from the general population in several aspects. Therefore, the present results should only be generalized to populations with a similar profile to that of the *Pró-Saúde* Study. However, analyses according to sex, schooling and occupation show similar performance across different strata, suggesting that ICPC-2 may also perform adequately in other contexts.

REFERENCES

1. Cohen J. A coefficient of agreement for nominal scales. *Educ Psych Meas*. 1960;20(1):37-46.
2. Faerstein E, Chor D, Lopes CS, Werneck GL. Estudo Pró-Saúde: características gerais e aspectos metodológicos. *Rev Bras Epidem*. 2005;8(4):454-66.
3. Lamberts H, Wood M. The birth of the International Classification of Primary Care (ICPC): Serendipity at the border of Lac Léman. *Fam Pract*. 2002;19(5):433-5.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-74.
5. Laurenti R. Décima revisão da Classificação Internacional de Doenças e Problemas Relacionados à Saúde (CID-10): a revisão do final do século. *Bol Oficina Sanit Panam*. 1995;118(3):273-80.
6. Lebrão ML, Carandina L, Magaldi C. Análise das condições de saúde e de vida da população urbana de Botucatu, São Paulo (Brasil). IV – Morbidade referida em entrevistas domiciliárias, 1983-1984. *Rev Saude Publica*. 1991;25(6):452-60.
7. Lebrão ML. Classificação internacional de motivos de consulta para assistência primária: teste em algumas áreas brasileiras. *Rev Saude Publica*. 1985;19(1):69-78.
8. Lebrão ML. Estudos de morbidade. São Paulo: Edusp; 1997.
9. Letrilliart L, Guiguet M, Flahault A. Reliability of report coding of hospital referrals in primary care versus practice-based coding. *Eur J Epidemiol*. 2000;16(7):653-9.
10. Reichenheim ME. Confidence intervals for the kappa statistic. *Stata J*. 2004;4(4):421-8.
11. Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res*. 1998;7(3):301-17.
12. Simões SEM, Reichenheim ME. Confiabilidade das informações de causa básica nas declarações de óbito por causas externas em menores de 18 anos no Município de Duque de Caxias, Rio de Janeiro, Brasil. *Cad Saude Publica*. 2001;17(3):521-31.
13. Szklo M, Nieto FJ. Quality assurance and control. In: Szklo M, Nieto FJ. Epidemiology: beyonds the basics. 2nd ed Jones and Bartlett: Boston; 2007. p.326-30.
14. Van der Heyden J, Van Oyen H, Gisle L. The classification of health problems in health interview surveys: using the International Classification of Primary Care (ICPC). *Soz Praventivmed*. 2004;49(2):161-3.
15. Comissão de Classificações da Organização Mundial de Ordens Nacionais, Academias e Associações Acadêmicas de Clínicos Gerais/Médicos de Família (WONCA). Classificação internacional de cuidados primários. 2. ed. Oxford: Oxford University Press; 1999.

Article based on the masters' dissertation by MMA Sampaio, presented to the Universidade Estadual do Rio de Janeiro, in 2006.
Financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – Edital Universal processes N. 471129/2003-8, 473746/2004-2, 471979/2003-1, 26/170.550/2004).
MMA Sampaio was supported by CNPq (Proc. No.130156/2004-3; Masters' scholarship).