REVISTA BRASILEIRA DE MEDICINA DE FAMÍLIA E COMUNIDADE
RBMFC

www.rbmfc.org.br

RESEARCH ARTICLES

Reliability and validity of ICPC-2-R encoding by medical students

Confiabilidade e validade da codificação CIAP-2 por estudantes de medicina

Confiabilidad y validez de la codificación CIAP-2 por estudiantes de medicina

Leonardo Ferreira Fontenelle
Álvaro Damiani Zamprogno
André Filipe Lucchi Rodrigues
Lorena Camillato Sirtoli
Natália Josiele Cerqueira Checon
Marcelo Santana Vetis
Diego José Brandão

Universidade Vila Velha (UVV). Vila Velha, ES, Brasil.
leonardof@leonardof.med.br *(Corresponding author)*;
alvarodamiani@hotmail.com;
andreflucchi@gmail.com;
lorenacamillatosirtoli@gmail.com;
nati_checon@hotmail.com;
marcelo.vetis@gmail.com;
diegojbrandao@yahoo.com.br

Abstract

Objective: To estimate how reliably and validly can medical students encode reasons for encounter and diagnoses using the International Classification of Primary Care, revised 2nd edition (ICPC-2-R). **Methods:** For every encounter they supervised during an entire semester, three family and community physician teachers entered the reasons for encounter and diagnoses in free text into a form. Two of four medical students and one teacher encoded each reason for encounter or diagnosis using the ICPC-2-R. In the beginning of the study, two three-hour workshops were held, until the teachers were confident the students were ready for the encoding. After all the reasons for encounter and the diagnoses had been independently encoded, the seven encoders resolved the definitive codes by consensus. We defined reliability as agreement between students and validity as their agreement with the definitive codes, and used Gwet's AC_1 to estimate this agreement. **Results:** After exclusion of encounters encoded before the last workshop, the sample consisted of 149 consecutive encounters, comprising 262 reasons for encounter and 226 diagnoses. The encoding had moderate to substantial reliability (AC_1, 0.805; 95% CI, 0.767–0.843) and substantial validity (AC_1, 0.864; 95% CI, 0.833–0.891). **Conclusion:** Medical students can encode reasons for encounter and diagnoses with the ICPC-2-R if they are adequately trained.

Keywords: Primary Health Care/classification; Observer Variation; Reproducibility of Results; Education, Medical, Undergraduate; Clinical Clerkship

Funding:
none declared.
Ethical approval:
1.688.334 (ES/UVV), approved on 07/28/2016
Conflict of interest:
none declared.
Provenance and peer review:
peer reviewed.
Received: 14/11/2017.
Accepted: 20/04/2018.

Cite as: Fontenelle LF, Zamprogno AD, Rodrigues AFL, Sirtoli LC, Checon NJC, Vetis MS, et al. Reliability and validity of ICPC-2-R encoding by medical students. Rev Bras Med Fam Comunidade. 2018;13(40):1-6. https://doi.org/10.5712/rbmfc13(40)1655

Resumo

Objetivo: Estimar a confiabilidade e a validade da codificação de motivos de consulta e problemas por estudantes utilizando a Classificação Internacional da Atenção Primária, 2ª edição (CIAP-2). **Métodos:** Para cada encontro supervisionado durante todo um semestre, três professores médicos de família e comunidade registraram os motivos de consulta e problemas em um questionário usando texto livre. Dois de quatro estudantes de medicina e um professor codificaram cada motivo de consulta ou problema usando a CIAP-2. No começo do estudo, houve duas seções de padronização com três horas de duração, até os professores julgarem que os estudantes estavam prontos para a codificação. Após todos os motivos de consulta e problemas terem sido codificados independentemente, os sete codificadores resolveram os códigos definitivos por consenso. Definiu-se confiabilidade como concordância entre estudantes, e validade como a concordância destes com os códigos definitivos; essa concordância foi estimada com o AC_1 de Gwet. **Resultados:** Após a exclusão dos encontros codificados antes da última sessão de padronização, a amostra consistiu em 149 encontros consecutivos, somando 262 motivos de consulta e 226 problemas. A codificação teve confiabilidade moderada a substancial (AC_1 0,805; IC 95% 0,767–0,843) e validade substancial (AC_1 0,864; IC 95% 0,833–0,891). **Conclusão:** Estudantes de medicina podem codificar motivos de consulta e problemas com a CIAP-2 se forem adequadamente treinados.

Palavras-chave: Atenção Primária à Saúde/classificação; Variações Dependentes do Observador; Reprodutibilidade dos Testes; Educação de Graduação em Medicina; Estágio Clínico

Resumen

Objetivo: Estimar la confiabilidad y la validez de la codificación de motivos de consulta y problemas de salud por estudiantes utilizando la Clasificación Internacional de Atención Primaria, 2ª edición (CIAP-2). **Métodos:** Para cada encuentro supervisado durante todo un semestre, tres profesores médicos de familia y comunidad registraron los motivos de consulta y los problemas de salud en un formulario usando texto libre. Dos de cuatro estudiantes de medicina y un profesor codificaron cada motivo de consulta o problema de salud utilizando la CIAP-2. En el comienzo del estudio, se llevaron a cabo dos sesiones de estandarización de tres horas, hasta que los profesores estuvieron seguros de que los estudiantes estaban listos para la codificación. Después de que todos los motivos de consulta y problemas de salud fueran codificados independientemente, los siete codificadores resolvieron los códigos definitivos por consenso. Se definió confiabilidad como concordancia entre los estudiantes y validez como la concordancia de éstos con los códigos definitivos; se estimó esta concordancia con el AC_1 de Gwet. **Resultados:** Después de la exclusión de los encuentros codificados antes de la última sesión de estandarización, la muestra consistió en 149 encuentros consecutivos, que comprendían 262 motivos de consulta y 226 problemas de salud. La codificación tuvo una confiabilidad moderada a sustancial (AC_1 0,805; IC 95% 0,767–0,843) y validez sustancial (AC_1 0,864; IC 95% 0,833–0,891). **Conclusión:** Los estudiantes de medicina pueden codificar los motivos de consulta y los problemas de salud con la CIAP-2 si fueran adecuadamente capacitados.

Palabras clave: Atención Primaria de Salud/clasificación; Variaciones Dependientes del Observador; Reproducibilidad de Resultados; Educación de Pregrado en Medicina; Prácticas Clínicas

Introduction

A member of the Word Health Organization's family of international classifications, the International Classification for Primary Care (ICPC) was designed to address the needs of family practice. Its rubrics were chosen to include only common reasons for encounter and diagnoses, occurring at least once per thousand patients-year. Furthermore, it allows for the description of morbidity in terms of episodes of care, as well as in terms of encounters.[1,2]

The ICPC is a biaxial classification system, with rubrics consisting of a one-letter code for the chapter followed by a two-digit numeric code which is part of a component. The chapters indicate the localization in a body system or as general, psychological or social, and the components indicate the code as a symptom or complaint (component 1), procedure (components 2-6) or diagnosis (component 7). While localization takes precedence over etiology in the chapters, the diagnosis component has etiological sub-components: infection, neoplasm, injury, congenital and other diagnoses.[2,3]

Although the reproducibility of ICPC encoding by medical doctors has been studied,[3-8] less is known about the reliability of the encoding by medical students. Medical students might need to use the ICPC during their clinical clerkship or internship in primary care, either to quantitatively report their activity to the medical school or to comply with the service routines (e. g. because of adoption in Brazil's public health system).[9]

Our objective was to estimate how reliably and validly can medical students encode reasons for encounter and diagnoses using the revised 2nd edition of the ICPC (ICPC-2-R).

Methods

The Vila Velha University (*Universidade Vila Velha* – UVV) Medical School is located in Vila Velha, one of the main municipalities of the Espírito Santo state, in Brazil. Following a guideline from the Brazilian Society of Family and Community Physicians (*Sociedade Brasileira de Medicina de Família e Comunicade* – SBMFC) and the Brazilian Association for Medical Education (*Associação Brasileira de Educação Médica* – ABEM),[10] UVV medical students participate in a Learning, Services and Community Interaction Program (*Programa de Interação Serviço, Ensino e Comunidade* – PISEC) during the four years before internship (which comprises the latter two years). The PISEC consists of practical activities in a primary care setting, beginning with an emphasis on territory and community in the two first years (PISEC 1 through 4) and focusing on family and clinical aspects of health in the next two years (PISEC 5 through 8).

The students are split in groups of approximately ten, each undertaking their activities in a different primary care center (*unidade básica de saúde* – UBS) under the supervision of a different teacher. In the next calendar semester, both the groups and their teachers stay in the same UBS, but every other semester the groups change teachers. While in PISEC 1 through 4 the teachers are mostly other health professionals, in PISEC 5 though 8 the teachers are primary care physicians.[11]

Three family and community physicians (LFF, MSV, DJB) are the teachers in a specific UBS for PISEC 5 through 8. During the second semester of 2016, every time a student discussed a consultation in this UBS with one of these teachers, the teacher recorded the reasons for encounter and diagnoses in free text into a form adapted from Gusso & Benseñor.[12] The form includes fields for age and gender, but does not identify the patient in any way. Thus, the data were analyzed per encounter, not per episode of care.

Two of four sixth-semester UVV medical students (ADZ, AFLR, LCS, NJCC) and one of the three teachers independently encoded each reason for encounter or diagnosis, using the Brazilian translation of the ICPC-2-R as the reference.[2] After data started being collected and encoded, the seven authors started holding three-hour workshops to compare their work. There were two workshops, after which the authors felt the four medical students were ready. After all data had been collected and encoded, definitive codes were decided by discussion among all seven authors. In this article, we analyze only the encoding done after the last workshop.

We defined reliability as the agreement between students, and validity as the agreement of students with the definitive codes. After describing them with percent observed agreement, we estimated reliability and validity using Gwet's first-order agreement coefficient (AC_1).[13] Compared to Cohen's κ, Gwet's AC_1 also ranges from -1 to +1, but is more resistant to known "paradoxes".[13,14] After estimating reliability and validity for reasons for encounter and for diagnoses separately, we combined them with inverse variance weighting into a single estimate for reliability and another single estimate for validity.

We expressed uncertainty around the estimates with bootstrap confidence intervals (CI) set with the bias-corrected and accelerated method (BC_a),[15] with 10,000 replicates. These replicates were obtained by resampling with replacement from the encounters (stratified by course semester), instead of from the reasons for encounter or diagnoses, to accommodate correlation of codes within encounters.

We analyzed the data using the R language and environment for statistical computing, version 3.4, with a custom script which provided AC_1 estimates numerically identical to the original implementation (but more quickly, mostly because it didn't calculate the variance).

The Human Research Ethics Committee of UVV approved the study (CAAE 57586516.9.0000.5064), and deemed written consent (by patients or students) unnecessary, because teachers collected only data that was routinely received, and the collected data was completely anonymous.

Results

During the study period there were 226 encounters for PISEC 5 through 8 in the specific UBS. After discarding the 77 encounters encoded before the last workshop, the sample consisted of 149 encounters, which comprised 262 reasons for encounter and 226 diagnoses. Median age was 31 years (range, 0–96), and 109 (73.2%) patients were female.

Students agreed among themselves in 78.6% of the reasons for encounter and 82.7% of the diagnoses, at the rubric level (Table 1). The corresponding chance-corrected coefficients were 0.785 (95% CI, 0.735–0.831) and 0.826 (95% CI, 0.775–0.871), with an overall reliability coefficient of 0.805 (95% CI, 0.767–0.843).

Table 1. Reliability of the ICPC-2-R encoding by medical students.

	Rubrics			Chapters		
	p_o	AC_1	95% CI	p_o	AC_1	95% CI
Reason for encounter	78.6%	0.785	0.735–0.831	93.1%	0.927	0.891–0.955
Diagnosis	82.7%	0.826	0.775–0.871	92.5%	0.920	0.878–0.951
Overall	80.6%	0.805	0.767–0.843	92.8%	0.924	0.894–0.948

AC_1, Gwet's first order agreement coefficient. ICPC-2-R, International Classification for Primary Care, revised 2nd edition. p_o, proportion of observed agreement.

Students agreed with the definitive code in 83.1% of the reasons for encounter and 89.3% of the diagnoses, at the rubric level (Table 2). The corresponding chance-corrected coefficients were 0.828 (95% CI, 0.786–0.866) and 0.892 (95% CI, 0.856–0.922), with an overall validity coefficient of 0.864 (95% CI, 0.833–0.891).

Table 2. Validity of the ICPC-2-R encoding by medical students.

	Rubrics			Chapters		
	p_o	AC_1	95% CI	p_o	AC_1	95% CI
Reason for encounter	83.1%	0.828	0.786–0.866	93.9%	0.935	0.905–0.958
Diagnosis	89.3%	0.892	0.856–0.922	95.2%	0.949	0.921–0.969
Overall	86.6%	0.864	0.833–0.891	94.4%	0.940	0.915–0.959

AC_1, Gwet's first order agreement coefficient. ICPC-2-R, International Classification for Primary Care, revised 2nd edition. p_o, proportion of observed agreement.

Discussion

Main findings

In Shrout's classification of agreement coefficients,[16] medical students can use the ICPC-2-R to encode reasons for encounter and diagnoses with substantial validity and moderate to substantial reliability. Both the reliability and the validity seemed to be moderately higher for the encoding of diagnoses than for the encoding of reasons for encounter. Obviously, the reliability and validity of the encoding at the chapter level were higher than at the rubric level.

Strengths and limitations

In this study, the standards for reliability and validity were relatively high: instead of a pair of student encoders, there were different six pairs among four student encoders; and agreement with the teacher was not always enough for the encoding to be considered valid. Furthermore, the sample size allowed for reasonably precise estimates, although not for subgroup analysis. Finally, using the block bootstrap resulted in theoretically more correct confidence intervals than the usual normal distribution-based confidence intervals.

Because the data were collected in a paper form, they were analyzed per encounter, instead of per episode of care. This was not a major issue, however, because this study didn't evaluate the correlation between reasons for encounter and diagnoses. In fact, collecting the data per encounter – anonymously – allowed the study to waive patient consent, thus eliminating losses by refusal and enhancing the generalizability of the findings.

As another potential limitation, the AC_1 confidence intervals don't technically generalize to encoders others than those participating in this study. However, we believe encoder training and previous abstraction of reasons for encounter and diagnoses are much more material to the generalization of our results than the potential difference in the width of the confidence intervals.

Comparison with the scientific literature

The reliability in this study was higher than or as high as that found in previous studies,[4,6,8] although one should keep in mind that students in this study were encoding data which had been abstracted specifically for this purpose, instead of abstracting data themselves.

As far as we know, this is the first study claiming to assess the validity of the ICPC-2-R encoding. There are, however, at least two studies comparing encoders with different levels of experience or information.[5,7] Again, the validity of the encoding in this study was higher than or as high as that found in those studies, partly because the students in this study were encoding previously abstracted data.

Implications for research and/or practice

If medical students receive adequate training, the health services where they are trained can trust them to use the ICPC-2-R to encode reasons for encounter and diagnoses reliably and validly. Likewise, teachers can rely on them to quantify their encounters using the ICPC-2 for educational reasons, and if they participate in research teams in clinical epidemiology the encoding can be delegated to them.

As a side note, in this study Gwet's AC_1 was very similar to the proportion of observed agreement (that is, without correction for chance agreement; see Tables 1 and 2), as well as Cohen's κ (data not shown). Because encoding with the ICPC-2-R means choosing among hundreds of different rubrics, chance agreement is negligible and thus there should be little difference between using Gwet's AC_1, Cohen's κ or even the proportion of observed agreement to assess reliability or validity of the ICPC-2-R encoding.

References

1. Soler JK, Okkes I, Wood M, Lamberts H. The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care. Fam Pract. 2008;25(4):312-7. https://doi.org/10.1093/fampra/cmn028
2. World Organization of National Colleges, Academies, and Academic Associations of General Practitioners/Family Physicians. Classificação Internacional de Atenção Primária (CIAP 2). 2ª ed. Florianópolis: Sociedade Brasileira de Medicina de Família e Comunidade; 2009.
3. Lamberts H, Wood M, Hofmans-Okkes IM. International primary care classifications: the effect of fifteen years of evolution. Fam Pract. 1992;9(3):330-9. https://doi.org/10.1093/fampra/9.3.330
4. Britt H, Angelis M, Harris E. The reliability and validity of doctor-recorded morbidity data in active data collection systems. Scand J Prim Health Care. 1998;16(1):50-5. https://doi.org/10.1080/028134398750003412
5. 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. https://doi.org/10.1023/A:1007609718223
6. Sampaio MM, Coeli CM, Miranda NN, Faerstein E, Werneck GL, Chor D, et al. Interobserver reliability of the International Classification of Primary Care. Rev Saúde Pública. 2008;42(3):536-41. https://doi.org/10.1590/S0034-89102008005000013
7. Sampaio MMA, Coeli CM, Alves MG, Soares MF, de Camargo KR Jr, Moreno AB. Confiabilidade interobservador da classificação internacional de atenção primária em uma unidade de atenção básica à saúde. Rev Bras Epidemiol. 2012;15(2):355-62. https://doi.org/10.1590/S1415-790X2012000200013
8. Frese T, Herrmann K, Bungert-Kahl P, Sandholzer H. Inter-rater reliability of the ICPC-2 in a German general practice setting. Swiss Med Wkly. 2012;142:w13621. https://doi.org/10.4414/smw.2012.13621
9. Basílio N, Ramos C, Figueira S, Pinto D. Worldwide Usage of International Classification of Primary Care. Rev Bras Med Fam Comunidade. 2016;11(38):1-9. https://doi.org/10.5712/rbmfc11(38)1225
10. Demarzo MMP, Almeida RCC de, Marins JJN, Trindade TG, Anderson MIP, Stein AT, et al. Diretrizes para o ensino na Atenção Primária à Saúde na graduação em Medicina. Rev Bras Med Fam Comunidade. 2011;6(19):145-50. https://doi.org/10.5712/rbmfc6(19)116
11. Dalla MDB, de Moura GAG, Bergamaschi MS. Metodologias ativas: um relato de experiência de estudantes de graduação em medicina da Universidade Vila Velha na disciplina de Interação Comunitária. Rev Bras Med Fam Comunidade. 2015;10(34):1-6. https://doi.org/10.5712/rbmfc10(34)647
12. Gusso GDF, Benseñor IM. A methodological proposal to research patients' demands and pre-test probabilities using paper forms in primary care settings. Rev Bras Med Fam Comunidade. 2013;8(27):97-105. https://doi.org/10.5712/rbmfc8(27)692
13. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(1):29-48. https://doi.org/10.1348/000711006X126600
14. Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9. https://doi.org/10.1016/0895-4356(90)90158-L
15. Efron B. Better Bootstrap Confidence Intervals. J Am Stat Assoc. 1987;82(397):171-85. https://doi.org/10.1080/01621459.1987.10478410
16. Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res. 1998;7(3):301-17. https://doi.org/10.1177/096228029800700306