Methods to Achieve High Interrater Reliability in Data Collection From Primary Care Medical Records

Clare Liddy, MD, MSc, CCFP, FCFP[1,2,3]

Miriam Wiens, BSc, MSc[2,3]

William Hogg, HonsBSc, MSc, MClSc, MD, CCFP, FCFP[1,2,3,4]

[1]Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada

[2]C. T. Lamont Primary Health Care Research Centre, Ottawa, Ontario, Canada

[3]Élisabeth Bruyère Research Institute, Ottawa, Ontario, Canada

[4]The Institute of Population Health, Ottawa, Ontario, Canada

MORE ONLINE
www.annfammed.org

Conflicts of interest: authors report none.

CORRESPONDING AUTHOR

Clare Liddy, MD, MSc, CCFP, FCFP
Élisabeth Bruyère Research Institute
43 Bruyère St
Ottawa, ON
Canada K1N 5C8
cliddy@bruyere.org

ABSTRACT

PURPOSE We assessed interrater reliability (IRR) of chart abstractors within a randomized trial of cardiovascular care in primary care. We report our findings, and outline issues and provide recommendations related to determining sample size, frequency of verification, and minimum thresholds for 2 measures of IRR: the κ statistic and percent agreement.

METHODS We designed a data quality monitoring procedure having 4 parts: use of standardized protocols and forms, extensive training, continuous monitoring of IRR, and a quality improvement feedback mechanism. Four abstractors checked a 5% sample of charts at 3 time points for a predefined set of indicators of the quality of care. We set our quality threshold for IRR at a κ of 0.75, a percent agreement of 95%, or both.

RESULTS Abstractors reabstracted a sample of charts in 16 of 27 primary care practices, checking a total of 132 charts with 38 indicators per chart. The overall κ across all items was 0.91 (95% confidence interval, 0.90-0.92) and the overall percent agreement was 94.3%, signifying excellent agreement between abstractors. We gave feedback to the abstractors to highlight items that had a κ of less than 0.70 or a percent agreement less than 95%. No practice had to have its charts abstracted again because of poor quality.

CONCLUSIONS A 5% sampling of charts for quality control using IRR analysis yielded κ and agreement levels that met or exceeded our quality thresholds. Using 3 time points during the chart audit phase allows for early quality control as well as ongoing quality monitoring. Our results can be used as a guide and benchmark for other medical chart review studies in primary care.

Ann Fam Med 2011;9:57-62. doi:10.1370/afm.1195.

INTRODUCTION

Despite advances in clinical information systems, patient chart audits are often the only way to collect required data for research.[1] Establishing rigorous methods for assessing the reliability (consistency) and validity (accuracy) of data is important.[2,3] Although there is evidence-based guidance on performing chart abstractions,[2,4-9] there is minimal and inconsistent advice for methods to ensure interrater reliability (IRR), such as selection of the sample size, frequency of reliability checks, and minimum thresholds for the κ statistic and percent agreement.[4,9-11]

There are currently no standard recommendations for the proportion of abstracted data that should be randomly checked for reliability,[12] and sample size calculations can yield dramatically different numbers.[5,13,14] Two methods commonly used in the literature, the goodness-of-fit approach[13] and the 95% confidence interval precision method,[5,15] rely on estimates that are difficult to determine without knowledge from a previous study. Additionally, few studies in the literature that use chart abstraction exam-

ine the reliability of the data critically[15,16] or justify the frequency and timing of reliability checks.

Finally, there have not been many IRR studies in primary care research,[5,6,17] to assist in determining the lowest threshold of data quality before repeated collection is required. A common interpretation of κ states that a value between 0.61 and 0.80 constitutes substantial agreement between raters, while in emergency medicine research, the benchmark is 95% agreement.[18] Ultimately, the published standards from primary care are few and have widely ranging values depending on the variables being measured.

We undertook this study to bridge some of the existing gap in reported methods for assessing IRR by describing a 4-part procedure for monitoring the quality of data collection[19] and how IRR was measured in a research study focused on quality improvement for cardiovascular care in Ontario, Canada.

METHODS

Study Design, Intervention, and Setting

The Improved Delivery of Cardiovascular Care (IDOCC) project is a practice-based study focused on implementing delivery system changes and evidence-based care in primary care (http://www.idocc.ca). The study is a cluster-randomized controlled trial that uses the stepped wedge design[20] to roll out the intervention over equally spaced time intervals. All primary care practices in the health region were eligible and invited to participate in a 24-month intervention. The main method of data collection is retrospective medical chart abstraction for a random selection of patients at high risk for cardiovascular disease. In phase I of the study, we recruited 27 practices (approximately 25% of those invited) and audited a mean of 62 charts per practice (SD = 10.77, range = 44-66). The Ottawa Hospital Research ethics board approved the study procedures.

Data Quality Monitoring Procedure

We developed a 4-part data quality monitoring procedure consisting of (1) use of a standardized protocol and standardized forms by chart abstractors; (2) extensive training in data abstraction; (3) continuous monitoring of κ values and percent agreement between abstractors; and (4) continuous quality improvement, including retraining, editing of standard protocols, and providing of feedback. Each part is described in detail below.

Development of Standardized Manual and Data Entry Protocol

We developed a detailed chart abstraction manual for cardiovascular process of care and clinical indicators (available online at http://www.annfammed.org/cgi/content/full/9/1/57/DC1) based on previous protocols.[21] To promote data validity and reliability, we included glossaries of synonyms and short forms for medical terminology, comment areas for abstractor notes, and tips for items less commonly documented or more difficult to find. For example, a common place to find smoking status is at the front of the paper chart or in the cumulative patient profile in the electronic health record.

We ensured standardized data entry methods through use of a secure software system provided by ClinicalAnalytics (CA) 4.0, a product of TrialStat Corporation (Ottawa, Ontario), which was tailored to the study. CA 4.0 is a data capture system, designed primarily for health researchers, that allows chart abstractors to enter data on a laptop while in the family practice. Information entered into TrialStat is tracked and changes are recorded in an audit trail. The program has logic models embedded in the entry fields, so illogical responses are not accepted and warnings prompt chart abstractors of fields left incomplete. The start-up cost was $8,000, with a maintenance cost of $7,500 per year. The cost of closing the secured site will be $2,500.

Abstractor Training

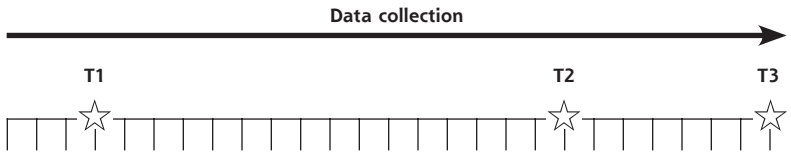
An experienced chart auditor trained 4 new auditors over 2 weeks with a standardized training program. In addition to a detailed review of the manual and chart audit exercises, typical scenarios of difficult data abstraction were presented.

We conducted a pilot phase in 5 practices where each auditor reviewed 5 medical charts under supervision of the trainer. Review of 5 charts was considered reasonable given existing time and resource constraints for participating practices. This process served as a pilot test for the data abstraction and electronic data collection forms, and addressed difficulties with data elements.

IRR Analysis and Continuous Quality Monitoring

We had 3 data quality checkpoints at which a different auditor reaudited a 5% sample of the charts and reliability was measured (Figure 1). A 5% reaudit matched the budgetary limits for phase I and was supported by the literature[4,22,23] but with no justification. The IRR checkpoints were unknown to chart abstractors and selected at random intervals. Timing was chosen to minimize practice disturbances and to fit within financial constraints and abstractor availability. The first time point was completed 1 week after abstraction began to ensure new chart abstractors had an early quality check. Additional quality checks would occur if major errors were detected or if staffing changes

Figure 1. Timeline of interrater reliability checks in phase I.



	T1	T2	T3
No. of practices finished with data collection at time point	4	20	27
No. of practices eligible for IRR check	4	16	7
No. of practices randomly selected	4	6	6
No. of chart abstractors being assessed	4	3	3
No. of charts randomly selected per practice	9	8	8
Total charts abstracted	36	48	48

IRR = interrater reliability; T1 = first check; T2 = second check; T3 = third check.

Notes: Vertical lines indicate number of practices whose charts were abstracted. Stars indicate when each check took place.

occurred during the 5-year project. The other 2 time points were dispersed within the data collection to ensure that new abstraction errors or patterns with different types of practices or charts were not introduced.

At the first IRR checkpoint, 9 charts were reabstracted in the first 4 practices that completed data collection, for a total of 14% (36 of 263) of charts completed. Each abstractor had performed chart abstraction at only 1 practice at this point, so practices were not randomly selected; however, charts within each practice were randomly selected.

We chose the number of charts between the other 2 reliability checkpoints to ensure consistent use of chart abstractor resources. The second IRR checkpoint thus occurred when one-half of the anticipated 40 practices had collected their data. We did not include the first 4 practices, which took part in the first IRR check, in this process since a sample of each auditor's charts had already been evaluated; hence, the second checkpoint sampled 16 practices instead of 20. We stratified the 16 practices by chart abstractor and randomly selected 2 practices per abstractor. Six practices (2 per abstractor as 1 abstractor no longer worked for the project) and 8 charts per practice were reaudited, for a total of 48 charts at the second time point.

The third IRR checkpoint occurred when all practices in the phase had completed data collection and used the methods described earlier. Because of lower practice recruitment, the final number of practices in phase I was 27; thus, the third IRR check occurred 1 month after the second. Two practices per chart abstractor were randomly selected from the 7 new practices, and 8 randomly selected charts per practice were reabstracted, for a total of 48 charts at the third time point. Chart abstractors performing the reabstraction were blinded to the responses of the first abstractor.

We used the Cohen κ statistic and percent agreement to assess the IRR; for both measures, higher values indicate greater IRR. We calculated the κ values for unbalanced 2×2 tables, using the method suggested by Gwet[24] and Crewson.[25]

Continuous Quality Improvement

We decided a priori that should IRR values be low, we would increase the number of IRR checkpoints. An overall κ (across all items) of 0.75 or less was the threshold set for abstractors to redo some or all of the data collection. We flagged individual items with a κ less than 0.70, a percent agreement less than 95%, or both, and incorporated them into post-IRR abstraction feedback and retraining sessions. The aim of these sessions was to highlight items for improvement and reasons for discrepancies, and to develop actions for correction. We retrained abstractors before they abstracted any more charts, giving feedback and solutions to common mistakes. In addition, the senior chart auditor was available during regular hours for telephone calls from abstractors. Questions that she could not address were forwarded to the principal investigators, who are also physicians.

RESULTS

The chart abstraction manual underwent several revisions. For example, we removed data collection fields for physical activity as this information was inconsistently recorded in the charts, and thus lacked validity and trustworthiness.

From the post-IRR abstraction feedback sessions, we developed 2 appendixes to the original chart abstraction manual. One appendix was a French acronym list for practices that had medical records mainly in French, and the other was the summary document with additional tips on where to find and how to record information.

The overall κ across all 38 items (diagnoses and quality of care indicators) measured was 0.91 (95% confidence interval, 0.90-0.92), and the overall percent agreement was 94.3%. Values for 23 items are shown in Table 1. For the sake of brevity, repetitive items (such as second and third blood pressure readings taken) and items for processes of care with low prevalence (such as referral for an electrocardiogram) are not shown in the table.

Of all 38 items measured, 10 had a κ value less than 0.75, and 3 of these were a value of 0.00 (range excluding zero values, 0.33-1.00). The 3 items with κ equal to zero also had high percent agreement (99%, 95%, and 99%) and unbalanced 2 × 2 tables where virtually all abstractors had indicated that the process of care did not occur. In such situations, the κ statistic will often overestimate or underestimate the agreement between 2 abstractors.[14,26,27] We therefore interpreted these items with caution and relied more on percent agreement as a guide to how these items were abstracted. Sixteen of the 38 items checked had less than 95% agreement (range, 87%-100%). Reaudit of medical record charts was not required at any point in phase I of the study.

Table 1. κ Statistics and Percent Agreement Values for Selected Data Abstraction Items (N = 132 Charts)

Item Abstracted	κ (95% CI)	Percent Agreement	
		Observed	Expected[a]
Diagnoses			
Cardiovascular disease	0.92 (0.85-0.99)	96.2	0.53
Chronic kidney disease	0.72 (0.58-0.87)	90.9	0.67
Diabetes	0.89 (0.82-0.97)	94.7	0.50
Dyslipidemia	0.90 (0.81-1.00)	97.0	0.69
Hypertension	0.80 (0.67-0.92)	93.2	0.66
Peripheral vascular disease	0.79 (0.62-0.97)	96.2	0.82
Stroke/TIA	0.89 (0.77-1.00)	97.7	0.79
Quality of care indicators[b]			
1 blood pressure recorded	1.00 (1.00-1.00)	100.0	0.87
1 HbA_{1c} test ordered	0.87 (0.78-0.96)	93.9	0.54
ASA recommended	0.83 (0.74-0.93)	91.7	0.33
CVD medications[c] recommended	0.85 (0.76-0.94)	92.4	0.50
eGFR test ordered	0.74 (0.62-0.87)	90.2	0.62
Fasting blood glucose test ordered	0.70 (0.56-0.84)	88.6	0.63
Glycemic control medications recommended	0.83 (0.75-0.92)	90.9	0.45
Hypertension medications[d] recommended	0.79 (0.68-0.90)	90.2	0.54
Lipid profile test ordered	0.71 (0.58-0.83)	87.1	0.56
Lipid-lowering medications recommended (statin or other)	0.79 (0.68-0.90)	90.9	0.57
Smoking status recorded	0.75 (0.63-0.88)	90.2	0.60
Smoking cessation counseling	0.88 (0.77-0.98)	96.2	0.69
Smoking cessation medication	0.88 (0.77-0.98)	96.2	0.70
Smoking cessation program referral	0.87 (0.77-0.98)	96.2	0.70
Waist circumference measured	0.00 (0.00-0.00)	95.5	0.95
Weight management program referral	0.86 (0.74-0.97)	95.5	0.68

ASA = acetylsalicylic acid (aspirin); CVD = cardiovascular disease; eGFR = estimated glomerular filtration rate; HbA_{1c} = hemoglobin A_{1c}; TIA = transient ischemic attack.

[a] Percent agreement expected (p_e) is a measure of the agreement that is expected to occur by chance between the raters.

[b] Measurement period was the 12 months before first date of abstraction.

[c] β-blocker, angiotensin-converting enzyme inhibitor, angiotensin receptor blocker.

[d] β-blocker, angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, diuretic, calcium channel blocker.

DISCUSSION

We designed a 4-part data quality monitoring method culminating in IRR analysis. We found that our data were of an acceptable level as defined by our a priori set standards. A 5% chart abstraction rate and IRR analysis showed a κ of 0.91 and percent agreement of 94.3%, signifying excellent agreement between abstractors. No charts needed to be reabstracted, supporting the effectiveness of our training and data collection approach.

When we developed our strategy in 2007, there was limited information on which to base our sample size. We opted for a sample size that was based on the literature and that fit within the study budget, with the intention of using the findings to inform future IRR analysis in other phases of the study.[5,27]

Although the Cohen κ is a commonly used statistic to measure IRR, there are known issues with its interpretation as other factors can influence the magnitude of the coefficient. The κ statistic can be influenced by the prevalence of the outcome of interest and potential bias between raters regarding the frequency of occurrence of the outcome; the κ in these situations may be higher or lower than the true chance-corrected measure of agreement.[14,26,28,29] We therefore chose to use percent agreement as an aid to interpreting κ, and to develop an overall sense of reliability. The data

on percent agreement for these items suggest that the actual reliability of the data was good and that our results where κ was very low (zero values) were an artifact of the manner in which κ was calculated. For these particular items (having a waist circumference measured in the last year, and in the year prior, and having weight control medications prescribed) the prevalence was low (they were not likely to occur in the practices); thus, the abstractors consistently chose 1 option. As a result, the 2 × 2 table used to calculate κ was unbalanced, and the expected agreement between abstractors was high, which produced κ values of 0.00.

The main objective of checking IRR is to obtain a sense of reliability and, thus, an indication of the quality of the data. A subjective element of IRR studies is the cut point at which a research team decides the data are not of high enough quality and, thus, medical chart audits must be repeated. Again, because descriptions in the primary care literature on the processes for ensuring data quality from primary care chart audits were limited, the research team had to arbitrarily select acceptable quality thresholds in this study, namely, a κ value of 0.75, a 95% agreement, or both. These cut points are consistent with the emergency medicine literature; however, hospital or ambulatory records would differ from those in primary care, particularly when it comes to standardized records and the type of data collected. As a result, these thresholds may not apply in primary care where records vary among clinics.

Eder et al[6] reported on procedures used to promote IRR in data abstraction for the quality of breast and cervical cancer screening services in California, and found high values of agreement, ranging from 89% to 100%, and a κ ranging from 0.59 to 1.0. Similarly, To et al[17] found an overall percent agreement of 93% and an overall κ of 0.81 when they examined IRR between the study chart abstractor and an experienced non-study chart abstractor, using 8 fictitious medical charts.

Quality of data obtained during a retrospective medical chart audit is limited to the availability and condition of the data housed within a patient's record.[4,16] Various sources of bias may affect the initial chart data, including the communication of ailments to a medical professional, followed by the entry of that information into a medical record. The quality of data is therefore influenced by whether the required information is available in a form that may be abstracted[4] and is accurate. Recent primary care reforms including incentive payments for prevention activities (eg, smoking cessation fees) and use of a team approach (emphasizing the need to communicate information to others) will also influence the data. Additionally, with a greater emphasis on electronic health systems that are tied to remuneration and performance (both within offices and within health regions), one could expect greater consistency for those targeted processes of care. Stange et al[30] suggest that given these external limitations on data quality, investigators should consider other methods that may be optimal for measuring a particular aspect of health care delivery. They demonstrated that methods to triangulate data sources were feasible within primary care practices and important when a reference standard for data collection does not exist. In our study, we have wide variation of record keeping, ranging from offices using electronic medical records (22%), flow sheets, and patient summaries, to those using basic paper-based records. Despite these variations, we have shown that in primary care, we can expect to see close to 95% agreement between raters.

In response to a lack of guidance in the literature, Engel et al[16] developed the Medical Record Review (MRR) Conduction Model, which describes and frames the process of chart abstraction. According to this model, data quality is affected by individual entities, including the investigator, abstraction manual, abstractor, data source, abstraction tool, and data quality analysis, in a cyclical process. Each entity influences and provides feedback to the other, which underscores the need to report IRR to guide future research. This interplay was highlighted in the development of our protocol for assessing reliability of chart audits in measuring quality improvement for cardiovascular care in primary care practices in Ontario, Canada.

Recommendations for the Future

We recommend that researchers report their justifications for IRR methods so that others can gauge their appropriateness and use the information for future research. Gow et al[31] evaluated clinical research studies published in 5 high-impact cardiology journals in 2005 and found that only 27% of the studies reported interobserver variability of measured variables. The authors suggest that this general lack of reporting is partially due to research methodology quality control being of low priority for researchers and journal editors. Others propose that restricted space in journals may explain why such description is omitted.[4]

We suggest that if κ is the statistic being used to measure IRR, the percent agreement be used to aid in interpretation. Alternatively, bias and prevalence indices, as outlined by Byrt et al,[26] could be calculated to help determine the amount to which the estimates are affected.[14,28] These indices provide estimates of the effect of the prevalence paradox (whereby a higher proportion of yes or no responses can produce lower κ values) and the bias paradox (whereby the difference between the rater's proportion of yes responses produces unbalanced marginal totals in the 2 × 2 table

and higher κ values) and can create a more accurate interpretation of κ. In this study, a 5% sampling of charts for quality control using IRR analysis yielded κ and agreement levels that met or exceeded our quality thresholds. We recommend that high quality be defined as a κ of 0.75, a 95% agreement, or both. Additionally, auditing charts at 3 time points during the chart audit phase allows for early quality control as well as ongoing quality monitoring.

This study has highlighted some areas that could benefit from further refinement in terms of IRR studies involving data collection in primary care practices for research. We have shown that a high level of IRR is possible in the primary care setting in Ontario, Canada. The high-quality data collected within the first phase of our study is attributable to the attention placed on developing the chart audit manual, the rigorous and standardized training protocols, and a continuous quality improvement process. Our results can be used as a guide and benchmark for other medical chart review studies in primary care.

To read or post commentaries in response to this article, see it online at http://www.annfammed.org/cgi/content/full/9/1/57.

Key words: Primary care; chart abstraction; quantitative methods: measurement issues/instrument development

Submitted September 17, 2009; submitted, revised, July 13, 2010; accepted July 26, 2010.

References

1. Schoen C, Osborn R, Huynh P, Doty M, Peugh J, Zapert K. *On the Front Lines of Care: Primary Care Doctors' Office Systems, Experiences, and Views in Seven Countries*. Washington, DC: The Commonwealth Fund; 2006.
2. Nagurney JT, Brown DF, Sane S, Weiner JB, Wang AC, Chang Y. The accuracy and completeness of data collected by prospective and retrospective methods. *Acad Emerg Med*. 2005;12(9):884-895.
3. Wu L, Ashton CM. Chart review. A need for reappraisal. *Eval Health Prof*. 1997;20(2):146-163.
4. Allison JJ, Wall TC, Spettell CM, et al. The art and science of chart review. *Jt Comm J Qual Improv*. 2000;26(3):115-136.
5. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. *Am J Manag Care*. 2002;8(9):787-793.
6. Eder C, Fullerton J, Benroth R, Lindsay SP. Pragmatic strategies that enhance the reliability of data abstracted from medical records. *Appl Nurs Res*. 2005;18(1):50-54.
7. Gearing RE, Mian IA, Barber J, Ickowicz A. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. *J Can Acad Child Adolesc Psychiatry*. 2006;15(3):126-134.
8. Panacek EA. Performing chart review studies. *Air Med J*. 2007;26(5): 206-210.
9. Yawn BP, Wollan P. Interrater reliability: completing the methods description in medical records review studies. *Am J Epidemiol*. 2005; 161(10):974-977.
10. Lowenstein SR. Medical record reviews in emergency medicine: the blessing and the curse. *Ann Emerg Med*. 2005;45(4):452-455.
11. Worster A, Bledsoe RD, Cleve P, Fernandes CM, Upadhye S, Eva K. Reassessing the methods of medical record review studies in emergency medicine research. *Ann Emerg Med*. 2005;45(4):448-451.
12. Worster A, Haines T. Advanced statistics: understanding medical record review (MRR) studies. *Acad Emerg Med*. 2004;11(2):187-192.
13. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat Med*. 1992;11(11):1511-1519.
14. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005; 85(3):257-268.
15. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? *Ann Emerg Med*. 1996;27(3):305-308.
16. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical Record Review Conduction Model for improving interrater reliability of abstracting medical-related information. *Eval Health Prof*. 2009;32(3):281-298.
17. To T, Estrabillo E, Wang C, Cicutto L. Examining intra-rater and inter-rater response agreement: a medical chart abstraction study of a community-based asthma care program. *BMC Med Res Methodol*. 2008;8:29.
18. Pan L, Fergusson D, Schweitzer I, Hebert PC. Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool. *J Clin Epidemiol*. 2005;58(9): 918-923.
19. Engel L, Henderson C, Colantonio A. Eleven steps to improving data collection: guidelines for a retrospective medical record review. *Occupational Ther Now*. 2008;10(1):17-20.
20. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol*. 2006;6:54.
21. Hogg W, Gyorfi-Dyke E, Johnston S, et al. Conducting chart audits in practice-based primary care research: a user's guide. *Can Fam Physician*. 2010;56(5):495-496.
22. Alberta Health and Wellness. *Ambulatory Care Re-abstraction Study: Executive Report*. Alberta, Canada: Alberta Health and Wellness; 2006.
23. Reisch LM, Fosse JS, Beverly K, et al. Training, quality assurance, and assessment of medical record abstraction in a multisite study. *Am J Epidemiol*. 2003;157(6):546-551.
24. Gwet K. *Computing Inter-rater Reliability with the SAS System*. Statistical Methods for Inter-rater Reliability Assessment [series]. Gaithersburg, MD: STATAXIS Consulting; 2002. Report 3.
25. Crewson PE. A correction for unbalanced kappa tables. In: *Proceedings of the 26th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc; 2001:1-3.
26. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423-429.
27. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 3rd ed. Hamilton, ON: B.C. Decker Inc; 2008.
28. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543-549.
29. Szklo M, Nieto FJ. *Quality Assurance and Control. Epidemiology Beyond the Basics*. Sudbury, MA: Jones and Bartlett; 2004:343-404.
30. Stange KC, Zyzanski SJ, Smith TF, et al. How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patients visits. *Med Care*. 1998;36(6):851-867.
31. Gow RM, Barrowman NJ, Lai L, Moher D. A review of five cardiology journals found that observer variability of measured variables was infrequently reported. *J Clin Epidemiol*. 2008;61(4):394-401.