Modeling Canadian Quality Control Test Program for Steroid Hormone Receptors in Breast Cancer: Diagnostic Accuracy Study

Teresa Pérez, PhD,[*] Nikita Makrestsov, MD, PhD,[†‡§] John Garatt, RT(cyto),[‡] Emina Torlakovic, MD, PhD,[‡§] C. Blake Gilks, MD,[†‡] and Susan Mallett, DPhil, DipStat Open[∥]

Abstract: The Canadian Immunohistochemistry Quality Control program monitors clinical laboratory performance for estrogen receptor and progesterone receptor tests used in breast cancer treatment management in Canada. Current methods assess sensitivity and specificity at each time point, compared with a reference standard. We investigate alternative performance analysis methods to enhance the quality assessment. We used 3 methods of analysis: meta-analysis of sensitivity and specificity of each laboratory across all time points; sensitivity and specificity at each time point for each laboratory; and fitting models for repeated measurements to examine differences between laboratories adjusted by test and time point. Results show 88 laboratories participated in quality control at up to 13 time points using typically 37 to 54 histology samples. In meta-analysis across all time points no laboratories have sensitivity or specificity below 80%. Current methods, presenting sensitivity and specificity separately for each run, result in wide 95% confidence intervals, typically spanning 15% to 30%. Models of a single diagnostic outcome demonstrated that 82% to 100% of laboratories had no difference to reference standard for estrogen receptor and 75% to 100% for progesterone receptor, with the exception of 1 progesterone receptor run. Laboratories with significant differences to reference standard identified with Generalized Estimating Equation modeling also have reduced performance by meta-analysis across all time points. The Canadian Immunohistochemistry Quality Control program has a good design, and with this modeling approach has sufficient precision to measure performance at each time point and allow laboratories with a significantly lower performance to be targeted for advice.

Key Words: breast cancer, estrogen receptor, progesterone receptor, meta-analysis, sensitivity and specificity

(*Appl Immunohistochem Mol Morphol* 2016;24:679–687)

Since 2009, the Canadian Immunohistochemistry Quality Control program (CIQC, http://www.CIQC.ca), endorsed by the Canadian Association of Pathologists, has established an immunohistochemistry (IHC) quality control program including estrogen receptor (ER) and progesterone receptor(PR) freely available to Canadian clinical IHC laboratories. Currently laboratories receive feedback at each proficiency testing assessment (PTA). This research explores using statistical modeling to provide more robust understanding of performance at each PTA and individual laboratory performance across time.

In breast cancer, pathology diagnostic results provide disease verification and biomarker predictive testing (ER, PR, and human epidermal receptor 2) for personalized targeting of adjuvant hormone therapy–based treatments.[1] For ER and PR tests, a false-positive test result could lead to a patient having no opportunity of treatment benefit but exposure to side effects, whereas a false-negative result may deny access to potentially life-saving treatment. ER and PR IHC tests are complex with diagnostic performance depending on preanalytical factors (tissue handling from surgery to processing), analytical methods (IHC staining methods), and postanalytical performance (interpretation, delivery of results).[2–5] Previous work has shown that most errors result from variability in IHC staining rather than interpretation.[6,7]

Quality assessment programs have been set up in the United States (College of American Pathologists Quality Improvement Program, http://www.cap.org), United Kingdom (UK-NEQAS, http://www.ukneqas.org.uk), Nordic countries (Nordic QC, http://www.nordiqc.org), Canada (Canadian Immunohistochemistry Quality

Received for publication May 4, 2015; accepted June 12, 2015.

From the [*]Department of Statistics and OR III, Complutense University of Madrid, Madrid, Spain; [†]Department of Pathology and Laboratory Medicine, University of British Columbia; [‡]Canadian Immunohistochemistry Quality Control (CIQC) and Canadian Association of Pathologists; [§]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; and [∥]School of Health and Population Sciences, University of Birmingham, Birmingham, West Midlands, UK.

Supported by grants from Spanish Ministry of Education and Science MTM2013-40778-R (T.P.) and a financial contribution from Health Canada, through the Canadian Partnership Against Cancer.

N.M., J.G., E.T., and C.B.G. are affiliated to the Canadian Immunohistochemistry Quality Control (CIQC). The authors declare no conflict of interest.

Reprints: Susan Mallett, DPhil, DipStat Open, School of Health and Population Sciences, University of Birmingham, Birmigham B15 2TT, West Midlands, UK (e-mail: s.mallett@bham.ac.uk).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.appliedimmunohist.com.

Control, http://www.CIQC.ca), Australia (Royal College of Pathologists of Australasia, http://www.rcpaqap.com.au), and elsewhere.

There are 2 main types of PTA program, both providing valuable information: (i) calibration of Class I IHC tests providing descriptive information on lineage or subclassification of lesion and (ii) level of concordance to reference standard across a sample of IHC tests (ie, diagnostic accuracy) for Class II IHC tests as recommended in guidelines for breast cancer IHC markers. Although some EQA programs (eg, UKNEQAS, NordiQC) continue to design PTA challenges for Class II IHC markers based on calibration, other programs (eg, CIQC, CAP) follow guidelines reporting results for breast cancer markers using concordance (accuracy). Both types of PTA are very valuable, but do not replace each other.

All PTA programs use unstained slides of formalin-fixed/paraffin-embedded tissue containing antigens of interest. In calibration PTA, tissues usually contain both positive cells (with range of antigen positivity) and negative cells. The number of tissue samples could vary, but even a single sample may be appropriate for calibration. There is a great value in comparing participating laboratory test calibration with a reference laboratory or consensus results.

In PTA designed to assess level of concordance (diagnostic accuracy), the design including the number of samples and statistical analysis needs to consider the statistical power, that is, the ability to make statistically robust recommendations in combination with practical considerations. In 2010, ASCO published the first official guidelines on levels of concordance for ER and PR testing,[1] recommending laboratories should achieve 45 correct results from 50 biomarker tests each for biomarker-positive and biomarker-negative breast cancers when introducing new tests. In addition, external quality assessment testing was recommended with 90% minimum concordance. Current diagnostic test accuracy for ER, PR, and human epidermal receptor 2 is estimated to include error rates of 10% to 20%[1,8] when comparing individual laboratories to reference laboratories.[1,9–11]

Statistically, a challenge for quality control programs measuring concordance is how to present test performance,[12] given limits on feasible numbers of cases per PTA. Monitoring clinical performance generalizable to clinical practice requires thresholds based on 95% confidence interval (CI) rather than an estimate of performance.

This research aims to enhance the quality of PTA for laboratories participating in the CIQC program, by investigating alternative analysis methods that overcome the limitations of giving feedback solely based on sensitivity and specificity from individual PTA.

MATERIALS AND METHODS

Study Population

Participating Laboratory Eligibility Criteria

All Canadian clinical IHC laboratories that provide testing for breast cancer markers ER and PR are eligible.

Setting

CIQC Program 2008 to 2012.

Clinical Samples

For 12 of 13 PTA, we assembled clinical samples from anonymized sequential clinical invasive breast carcinoma surgical cases (mastectomies, lumpectomies) from 490 cases from daily surgical pathology practice, aiming to represent full clinical spectrum of breast cancer cases by grade and histology, with tumors >5 mm. We used samples from reference laboratories with established internal preanalytical quality control (tissue procurement, fixation, and processing) assembled into paraffin block TMA of 37 to 54 samples, with single tissue cores 0.6 mm in diameter each containing 100 to 500 tumor cells, using manual tissue arrayer MTA1 (Beecher Instruments Inc, Silver Spring, MA). Same source tissues were used for ER and PR. Within each round, all laboratories received slides from the same cases. For run 12, duplicate samples of formalin-fixed paraffin-embedded cell blocks of breast carcinoma cell lines were used. Runs 6 and 11 are replica TMAs of the same cases.

Laboratories Methods

Participating laboratories used and submitted in-house validated protocols for IHC ER and PR assays varying in primary and secondary antibodies, antigen retrieval methods, and visualization chromogen enhancers (http://cpqa.ca).

Index test

Index test: CIQC sent serially cut unstained TMA slides to each laboratory, 1 slide per biomarker. Laboratories stain slides and return with local interpretation conforming to ASCO2010 guidelines (positive tests >1% positive tumor cells nuclei of any intensity).[1] In addition, results were assessed centrally by CIQC expert panel. Test results were classified as positive, negative, or uninterpretable. Uninterpretable defined when tissue section cannot be classified as positive or negative, including missing sample due to detachment, damage, or no tumor/insufficient tumor for biomarker evaluation, that is, generally fewer than 50 cancer cells.

Reference Standard

Benchmark consensus result using independent staining and interpretation blinded to other test interpretations, as a surrogate ER and PR reference, from group of CIQC "reference laboratories."[6]

Statistical Analysis

We initially performed a descriptive analysis of the overall sensitivity and specificity for each laboratory, using analysis across all laboratories and time points with simple pooling, as if all results from single study.

For ER and PR separately, we used meta-analysis for each laboratory across all time points, using Stata metandi[13] bivariate meta-analysis of sensitivity and specificity using complete case analysis.[14] We used xtmelogit[15] when metandi did not converge. Binary outcome data are typically modeled with logistic regression. We used conditional logistic regression using Generalized Estimating Equations method (GEE)[16] extensions of generalized linear models to accommodate correlated data. We used alternative methods of estimation[17] in to account for missing data. We examined differences between laboratories adjusted by test, time point, and also possible interactions between them. With this approach, the relationship among the outcome variable and predictor variables is estimated using all available data, including data from individuals with missing observations.

A multilevel, longitudinal model was fitted using GEE package in R[18]; multilevel clustering of individual cases across laboratories within each time point, longitudinal data from laboratories at up to 13 time points. All independent variables in model are qualitative and incorporated as nominal or numerical, test at 2 levels (ER and PR tests), laboratory with 78 levels (77 different laboratories participating in both tests and the reference test) and 13 time points. The result for each laboratory for each test is the dichotomous dependent variable (positive and negative outcomes). Statistical significance was set at $P \leq 0.05$.

RESULTS

A total of 88 laboratories participated according to CIQC workflow protocols (Fig. 1), taking part in up to 13 consecutive rounds of assessment (Table 1). Initially, 18 laboratories participated increasing to 74 in 2012. Participation is voluntary and most laboratories have frequent participation following entry to the program.

Presence and Issues of Uninterpretable Results

ER and PR IHC stained slide sections are reported as positive, negative, or uninterpretable. The percentage of uninterpretable or missing results varied across assessment runs (Supplementary Figure 1, Supplemental Digital Content 1, http://links.lww.com/AIMM/A86). Across all runs 16% of results are uninterpretable, and treated as missing data in analyses. The presence of uninterpretable results creates problems with analysis methods that ignore these missing results, such as simplest methods to calculate sensitivity and specificity. This was a key issue we considered when designing appropriate analysis methods.

There was a noticeable correlation between the percentage uninterpretable results in the reference standard and the laboratories, suggesting association with the assessment run, rather than individual laboratories. On the basis of the increase in uninterpretable results as the number of participating laboratories increases, a likely cause is an insufficient cancer cells per sample for valid interpretation, as a larger number of sections are cut from the same TMA block. Supporting this, there are no uninterpretable data in the reference for run 12 which uses cell culture lines.

Descriptive Analysis of Diagnostic Performance

A descriptive analysis of ER and PR sensitivity and specificity averaged by simple pooling across all assessment runs and samples for each laboratory is summarized (Supplementary Table 1, Supplemental Digital Content 2, http://links.lww.com/AIMM/A87). The percentage of laboratories with values of sensitivity and specificity above 95% are higher with ER test than with PR test (62% and 44% vs. 51% and 40%, respectively).

Supplementary Figure 2 (Supplemental Digital Content 3, http://links.lww.com/AIMM/A88) shows results for sensitivity and specificity for each laboratory based on meta-analysis across all times. Both sensitivity and specificity are higher than 90% (dotted lines) for almost all laboratories with both ER and PR, with no values below 80%. Interpretation of 95% CI is complex as it depends on both the number of PTA for each laboratory and accuracy.

Limitations of Measuring Performance Using Simple Methods at Each Time Point

Ideally, we would assess laboratory performance at each PTA, to compare laboratories and to monitor each laboratory over time. However, there are major limitations with using simple methods such as sensitivity and specificity at each time point, most importantly the maximum number of cases feasible per PTA, and the presence of uninterpretable data, which if ignored leads to overestimation of test performance. For monitoring of ER and PR, CIQC uses more cases at each time point than most other quality-monitoring programs.

We show sensitivity and specificity for ER at individual times for 1 laboratory in Figure 2, with the number of cases. The CIQC includes typically 37 to 54 samples per biomarker; sensitivity is calculated from reference-positive samples (median 27 ER+, 20 PR+) and specificity is calculated from reference-negative samples (median 12 ER− to 16 PR−). The CIQC PTA prevalence of ER+ and/or PR+ cases (range, 61% to 88% ER+; 45% to 70% PR+) reflects clinical practice, except at time point 12 which uses formalin-fixed paraffin-embedded cell blocks of breast carcinoma cell lines.

In Figure 2, the smaller sample size for specificity results in wider 95% CI, giving an unacceptable level of uncertainty for the estimate of specificity when applied to normal clinical practice. The 95% CI for sensitivity are narrower, but more ER-positive cases would be needed to assure the performance would be within acceptable limits in normal clinical practice.

The large number of tissue samples needed in each PTA for sufficiently precise 95% CI is beyond what is considered practically feasible. To follow ASCO guidelines on ER and PR hormone testing, a sample size of 144 biomarker-positive cases would be necessary to ensure a lower limit of 90% for sensitivity (ie, 90% as lower limit for 95% CI) based on 95% average sensitivity. Similarly for specificity, 144 biomarker-negative samples would be needed.

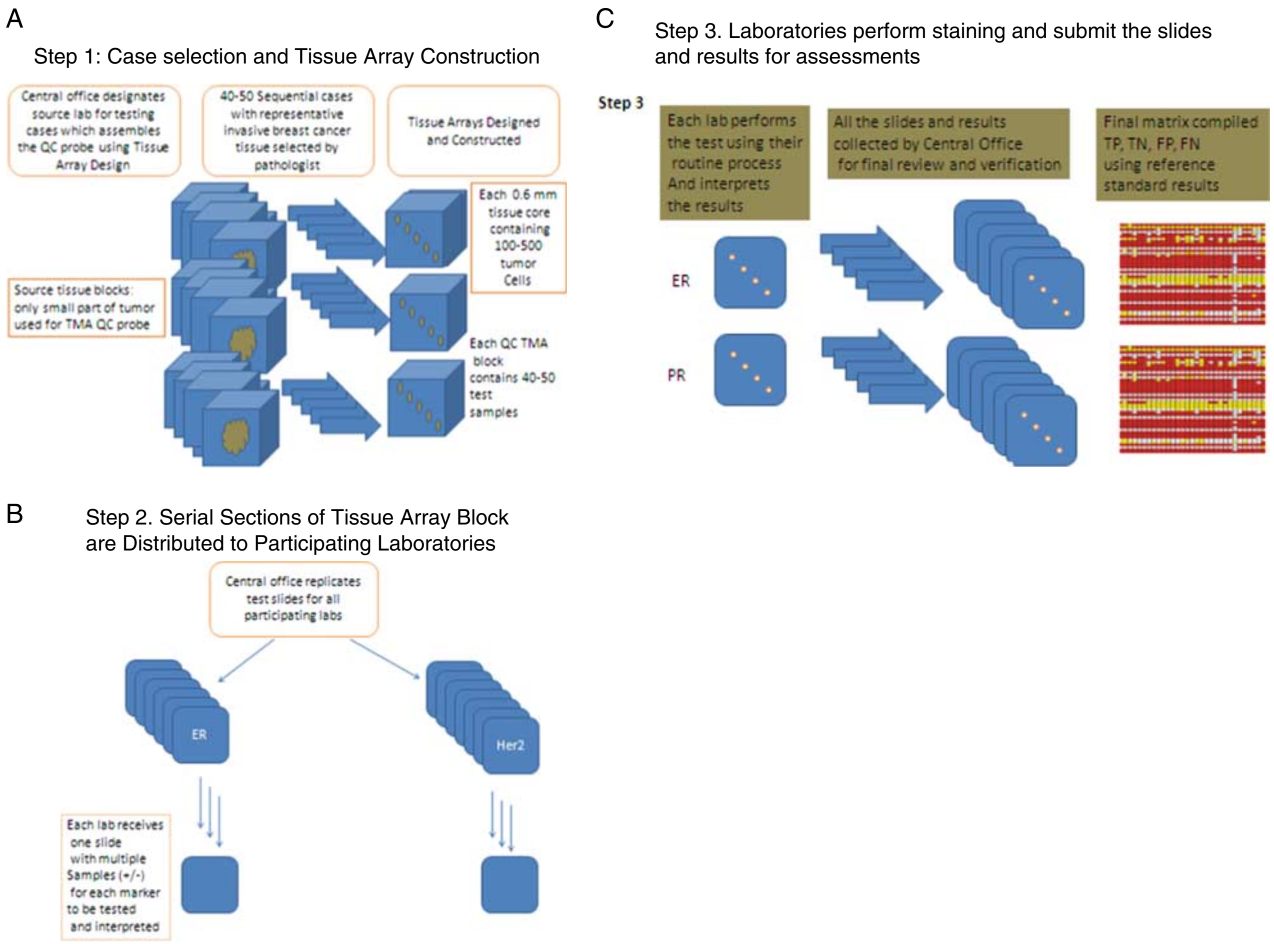
A
Step 1: Case selection and Tissue Array Construction
Central office designates source lab for testing cases which assembles the QC probe using Tissue Array Design
40-50 Sequential cases with representative invasive breast cancer tissue selected by pathologist
Tissue Arrays Designed and Constructed
Each 0.6 mm tissue core containing 100-500 tumor Cells
Source tissue blocks: only small part of tumor used for TMA QC probe
Each QC TMA block contains 40-50 test samples
C
Step 3. Laboratories perform staining and submit the slides and results for assessments
Step 3
Each lab performs the test using their routine process And interprets the results
All the slides and results collected by Central Office for final review and verification
Final matrix compiled TP, TN, FP, FN using reference standard results
ER
PR
B
Step 2. Serial Sections of Tissue Array Block are Distributed to Participating Laboratories
Central office replicates test slides for all participating labs
ER
Her2
Each lab receives one slide with multiple Samples (+/-) for each marker to be tested and interpreted

FIGURE 1. Project workflow. (A) Step 1: case selection and tissue array construction. (B) Step 2: serial sections of tissue array block are distributed to participating laboratories. (C) Step 3: laboratories perform staining and submit the slides and results for assessments.

TABLE 1. Number of Cases and Participating Laboratories at Each Proficiency Testing Assessment Time Point

Proficiency Testing Assessment Time Point	No. Participating Laboratories	No. Cases	No. ER Cases* (Negative, Positive)	No. PR Cases* (Negative, Positive)
1	18	37	12, 25	16, 20
2	23	37	14, 22	18, 15
3	25	54	11, 42	20, 29
4	32	44	11, 28	18, 20
5	57	40	13, 27	12, 20
6	59	40†	14, 26	12, 28
7	61	43	8, 31	13, 26
8	59	54	14, 27	17, 16
9	67	46	5, 36	13, 28
10	70	40	12, 19	17, 15
11	72	40†	12, 25	16, 20
12	71	18‡	12, 6	12, 6
13	74	46	10, 35	14, 31
Total across all assessments	88	490	136, 324§	182, 254§

*By reference standard.

†Test samples are the same in assessments 6 and 11, so are only counted once in totals.

‡Nine different test samples evaluated twice.

§Note that some cases have uninterpretable reference standard results, so the totals of negative and positive cases are lower than total cases. See Figure 2, where uninterpretable results are shown as missing data.

ER indicates estrogen receptor; PR, progesterone receptor.

Time	TP	FP	FN	TN	missing data	Sensitivity (95% CI)	Specificity (95% CI)
1	25	1	0	11	0	1.00 [0.86, 1.00]	0.92 [0.62, 1.00]
2	21	1	1	12	1	0.95 [0.77, 1.00]	0.92 [0.64, 1.00]
3	42	1	0	10	1	1.00 [0.92, 1.00]	0.91 [0.59, 1.00]
4	26	0	0	9	9	1.00 [0.87, 1.00]	1.00 [0.66, 1.00]
5	27	0	0	13	0	1.00 [0.87, 1.00]	1.00 [0.75, 1.00]
6	26	1	0	13	0	1.00 [0.87, 1.00]	0.93 [0.66, 1.00]
7	30	1	0	7	5	1.00 [0.88, 1.00]	0.88 [0.47, 1.00]
8	24	0	0	11	11	1.00 [0.86, 1.00]	1.00 [0.72, 1.00]
9	36	0	0	5	3	1.00 [0.90, 1.00]	1.00 [0.48, 1.00]
10	17	0	2	12	8	0.89 [0.67, 0.99]	1.00 [0.74, 1.00]
11	29	0	0	10	1	1.00 [0.88, 1.00]	1.00 [0.69, 1.00]
12	6	0	0	11	1	1.00 [0.54, 1.00]	1.00 [0.72, 1.00]
13	35	0	0	10	1	1.00 [0.90, 1.00]	1.00 [0.69, 1.00]

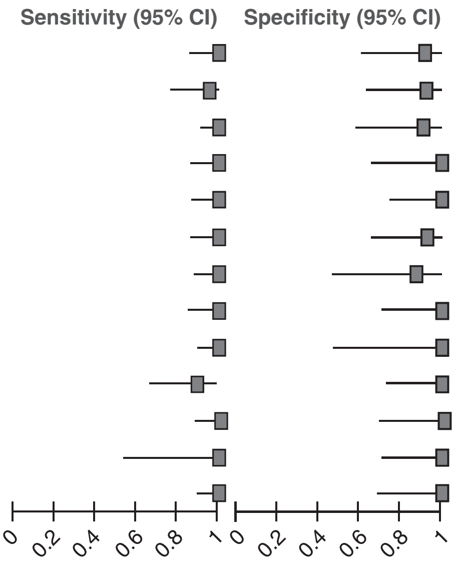

FIGURE 2. Sensitivity, specificity, and 95% CI for laboratory A with ER test by time points. CI indicates confidence interval; ER, estrogen receptorFN, false-negative; FP, false-positive; TN, true-negative; TP, true-positive.

Multilevel Longitudinal Statistical Modeling

Recognizing the limitations of the number of feasible cases in each assessment and the unavoidable presence of uninterpretable results, we sought to develop alternative performance analysis methods to enhance the quality assessment using the current program design.

We investigated an alternative analysis method to examine performance from individual PTA by including data from all time points and all laboratories together in a multilevel longitudinal regression model. We fitted a GEE model comparing ER and PR tests with the reference standard, allowing for data correlation within laboratories over time and including interactions (laboratories with test, laboratories with time). Interaction effects were statistically significant (Supplementary Table 2, Supplemental Digital Content 4, http://links.lww.com/AIMM/A89). We modeled ER and PR separately, to simplify interpretation of model coefficients.

Figure 3 presents results for laboratories which do not have significant differences between either PR or ER and the reference standard, that is, the 95% CI includes zero. Values closer to zero indicate better agreement with reference. Laboratories A and B present good results with both PR and ER (values close to zero for all time points), laboratories C and D perform better with the ER, and laboratories E and F have better results for the PR. It is important to realize that when the CIs are very narrow, there is low variability between individual cases indicating that cases with both positive and negative test results closely match the reference.

However, some laboratories show significant differences with the reference, with CIs not including zero and values far from zero. In Figure 4, we show examples of laboratories with significant differences to the reference at ≥ 3 time points.

Negative values mean the rate of false-negative results is higher, that is, IHC understaining causes significantly more false-negative results, compared with reference. Similarly, positive values mean that the rate of false-positive results is higher, that is, IHC overstaining causes significantly more false-positive results, than the reference standard. For example, results for the ER and/or PR tests of the laboratories G, J, K, and L tend to be more negative (higher rate of false-negative results, ie, lower sensitivity) than the reference. Laboratories H, I, and M have both positive and negative values. These laboratories at some time points undercall positive cases (false-negative results leading to lower sensitivity) but at other time points overcall positive results (false-positive results leading to lower specificity). CIs are wider (Fig. 4 vs. Fig. 3) due to the high variability between individual cases and reference.

At most time points, most laboratories show no significant difference of ER and PR results to the reference (82% to 100% of laboratories for ER, 75% to 100% for PR, except 1 PR run at 63%; Supplementary Table 3, Supplemental Digital Content 5, http://links.lww.com/AIMM/A90 shows percentage laboratories with a significant difference).

Comparison of Methods

We have compared results obtained with 2 more advanced statistical approaches, meta-analysis and GEE model. Figure 5 shows sensitivity and specificity from meta-analysis across all time points, for laboratories G to M identified with significant differences (Fig. 4).

Laboratories with mainly negative values, like J, K, and L with ER test and G and J with PR (Fig. 4), show below-average sensitivity (Fig. 5). Laboratories with higher positive coefficients, H and I with ER test and I

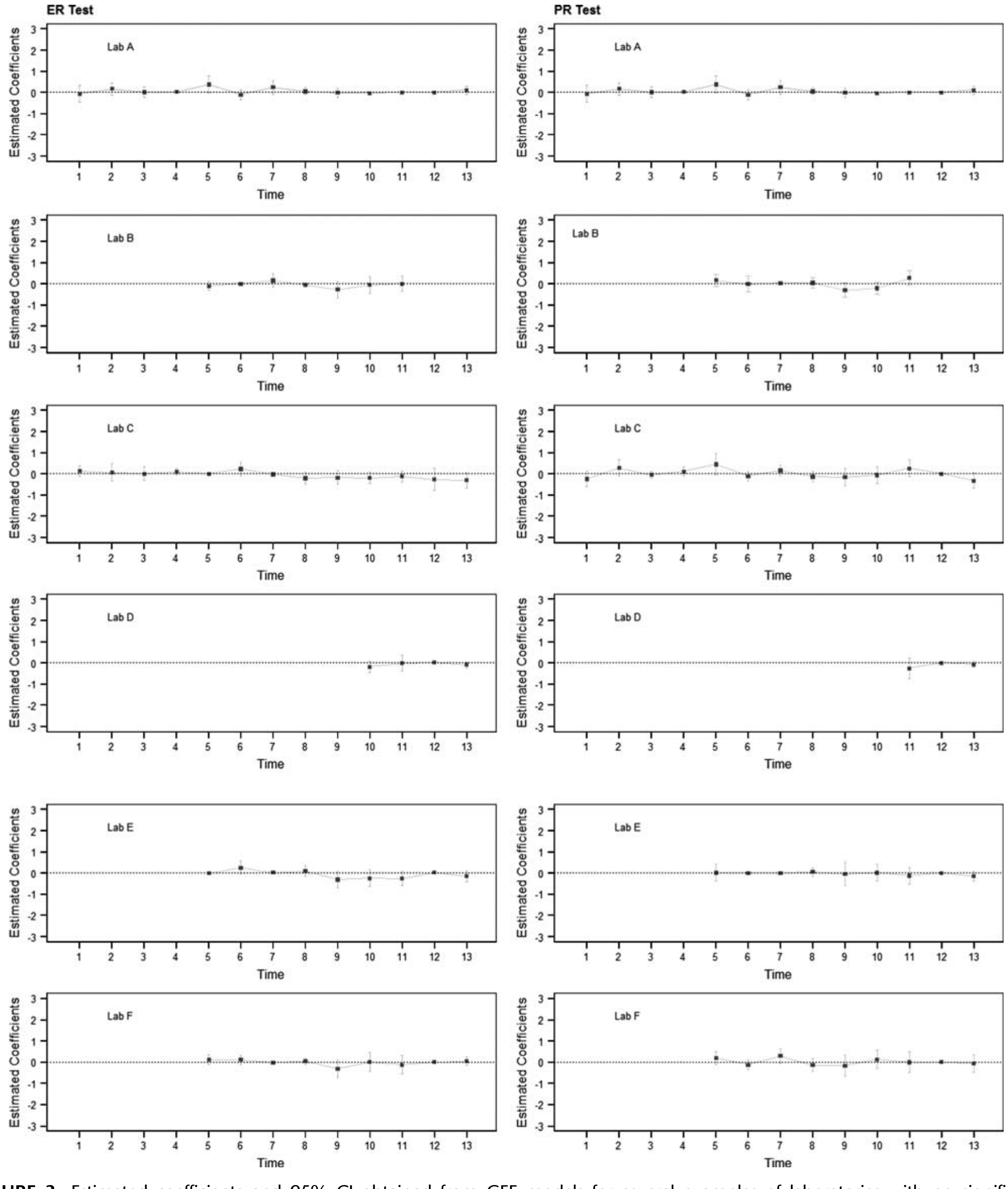

FIGURE 3. Estimated coefficients and 95% CI obtained from GEE models for several examples of laboratories with no significant differences to the reference standard. Each graph shows results from a single laboratory, with connected dots corresponding to PTA times where the laboratory participated. 95% CI are shown for each time point; where 95% CI overlap to the zero line (same accuracy as reference standard result), there is no significant difference between laboratory and reference standard results at that time point. Positive values indicate a higher rate of positive results than the reference standard, that is, false-positive results caused by IHC overstaining. Negative values indicate a higher rate of negative results compared with reference standard, that is, false-negative results caused by IHC understaining. Results for ER and PR tests are shown on the left and right side, respectively. CI indicates confidence interval; ER, estrogen receptor; GEE, Generalized Estimating Equations; IHC, immunohistochemistry; PR, progesterone receptor; PTA, proficiency testing assessment.

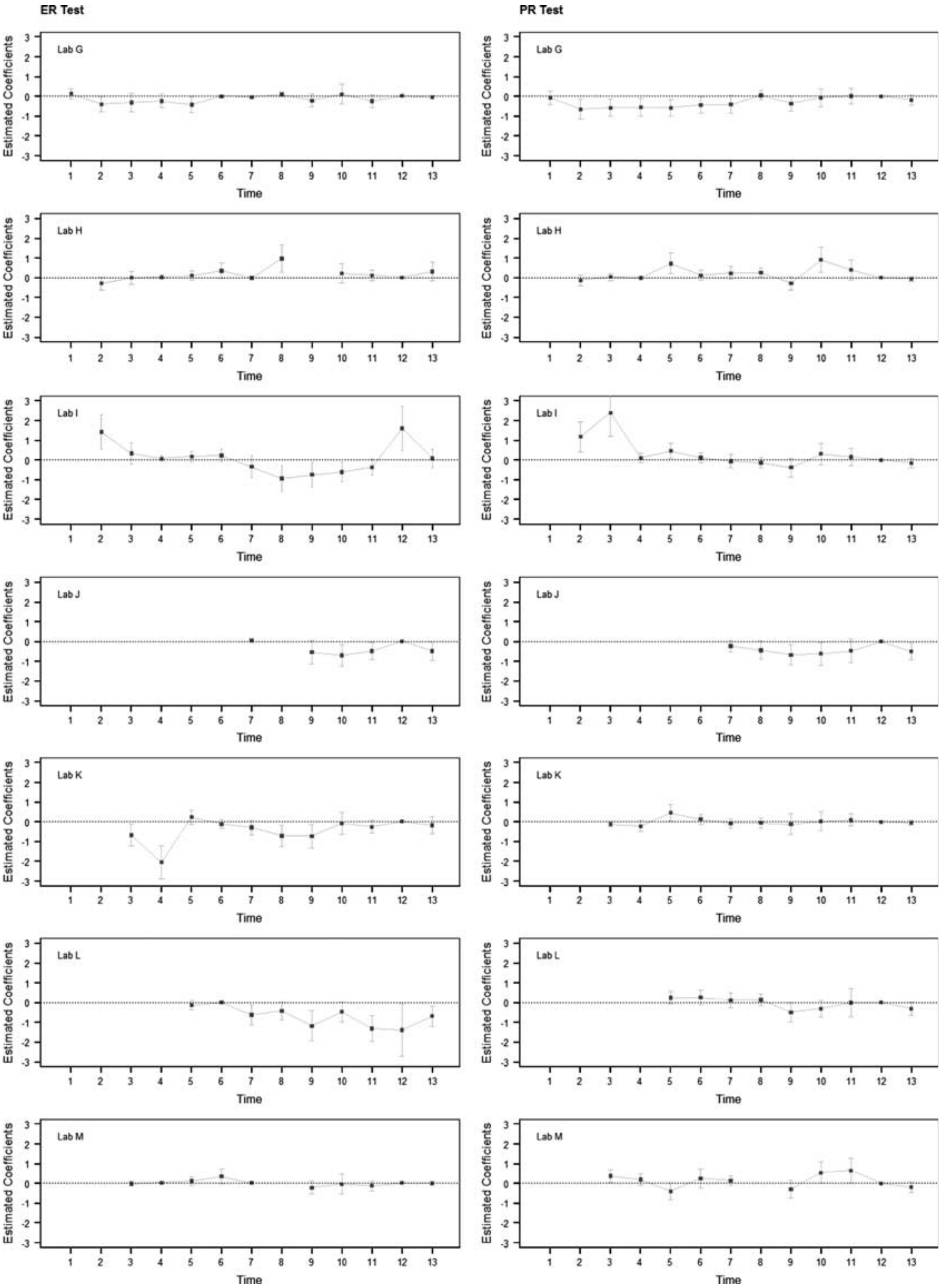

FIGURE 4. Estimated coefficients and 95% CI obtained from GEE models for some examples of laboratories with ≥ 3 PTA results with significant differences to the reference standard. Each graph shows results from a single laboratory, with connected dots corresponding to PTA times where the laboratory participated. 95% CI are shown for each time point; where 95% CI do not overlap to the zero line (same accuracy as reference standard results), there is a significant different between laboratory and reference standard results at that time point. Positive values indicate a higher rate of positive results than the reference standard, that is, false-positive results caused by IHC overstaining. Negative values indicate a higher rate of negative results compared with reference standard, that is, false-negative results caused by IHC understaining. Results for ER and PR tests are shown on the left and right side, respectively. CI indicates confidence interval; ER, estrogen receptor; GEE, Generalized Estimating Equations; IHC, immunohistochemistry; PR, progesterone receptor; PTA, proficiency testing assessment.

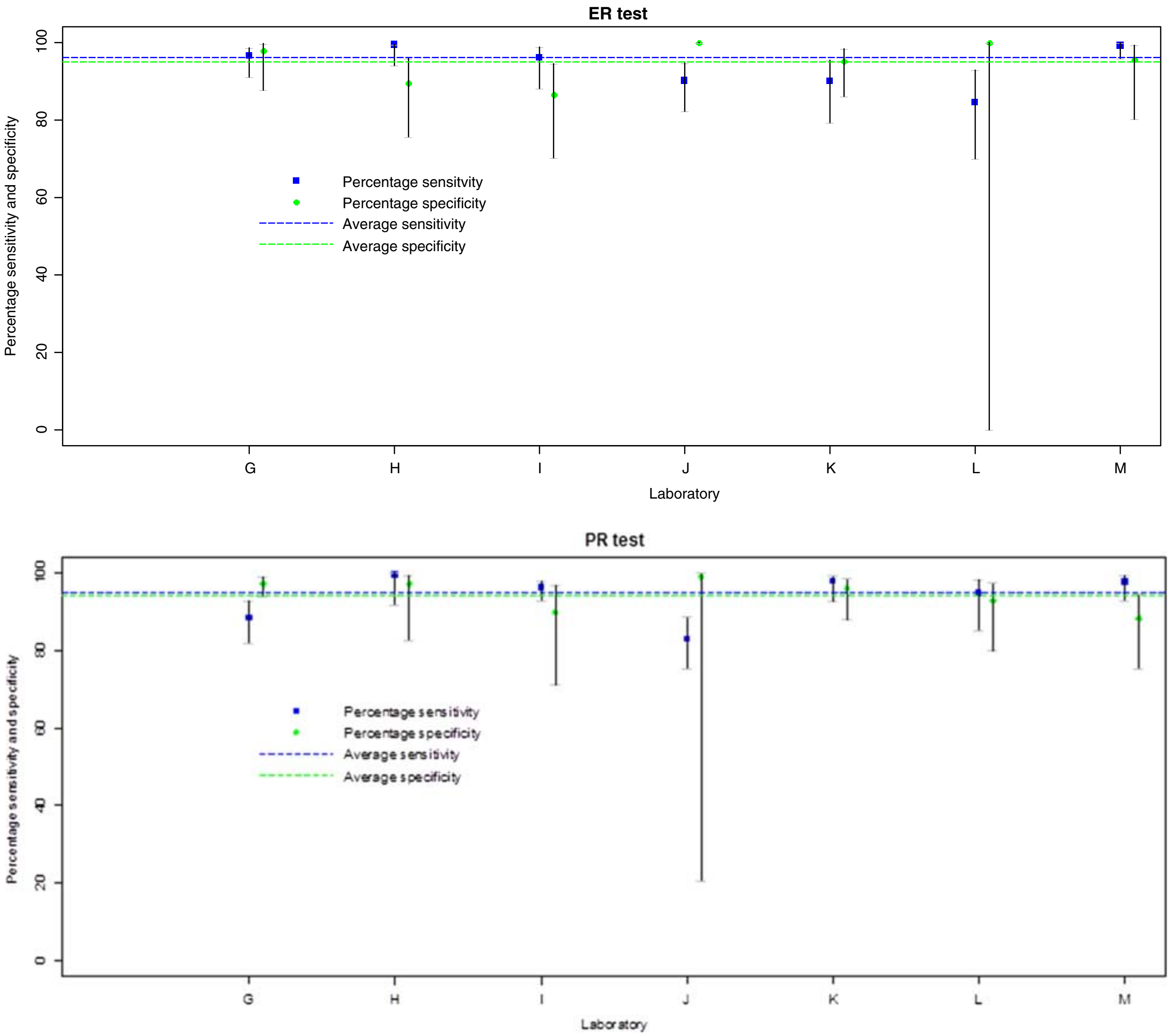

FIGURE 5. Percentage sensitivity, specificity, and 95% CI obtained in the meta-analysis for some examples of laboratories with ≥ 3 significant differences with reference standard according to GEE models for ER and PR tests. Meta-analysis results are shown for laboratories across all time points, providing less detailed information than GEE at separate time points. Laboratories with positive coefficients by GEE (H and I for ER) have lower values of specificity. Laboratories with negative coefficients by GEE have lower values of sensitivity. Separate graphs are shown for ER and PR. CI indicates confidence interval; ER, estrogen receptor; GEE, Generalized Estimating Equations; IHC, immunohistochemistry; PR, progesterone receptor.

and M with PR test (Fig. 4), have below-average specificity in Figure 5.

Comparison shows both methods identify the same performance issues in the same laboratories; however, the GEE method has an advantage that it allows the performance of laboratories to be monitored at each time point, whereas meta-analysis results are based on averages across all time points.

The GEE method also includes data from cases with missing observations, in contrast to the meta-analysis of sensitivity and specificity, where missing data are excluded resulting in overestimation of sensitivity and specificity.

DISCUSSION

This research investigates statistical modeling methods to analyze diagnostic performance of ER and PR IHC tests in CIQC participating laboratories. We compare laboratories to a reference standard and can distinguish those with satisfactory or different performance. Using GEE analysis, graphical display for individual laboratories over time allows understanding of consistency and direction (IHC overstaining or understaining) for test performance. Laboratories we identify using our accuracy coefficients as having significantly reduced test performance correspond to laboratories identified as unsatisfactory based on meta-analysis of sensitivity and specificity results across all time points. Our statistical modeling aids interpretation of laboratory performance because we model diagnostic accuracy with good precision (indicated by narrow width of 95% CI), allowing both robust comparison at a PTA monitoring over time.

ASCO 2010 guidelines on ER and PR hormone receptor testing[1] specify external quality assessment testing with at least 90% concordance for sensitivity and specificity, but do not give guidance on acceptable 95% CIs for generalizability of audit performance to normal clinical performance. No current monitoring programs have a sufficiently large number of cases at PTA to determine whether decreases in sensitivity or specificity are due to chance or to a real underlying difference in performance.

The CIQC program, where laboratories participate regularly and each assessment round includes typically 37 to 50 samples, allows fitting of GEE models. We achieve good precision by measuring performance using a single coefficient (instead of both sensitivity and specificity) and by appropriate analysis (clustering data from same patient samples, tests, laboratories, and PTA). The precision of our accuracy measure is sufficient to distinguish real (significant) differences between laboratories as opposed to statistical sampling error, which will avoid many overcalls or undercalls of unsatisfactory performance caused by variation between PTA. There is unavoidable variation between PTAs based on the number of: "difficult" patient samples, for example, lower biomarker expression resulting in lower intensity staining; tumor cells per slide affecting the chance of identifying positive cells; and ER-positive and PR-positive cases. We note there are alternative methods using a combined error rate, but these do not indicate the direction of errors (overstaining or understaining). Another advantage of our modeling methods is to account for missing data, although assumed as missing at random. In contrast, methods using sensitivity and specificity typically ignore missing data, causing overoptimistic estimates of performance.

Conclusions and Implications for Practice, Policy and Future Research

Assessment programs perform a vital role in providing quality control and monitoring diagnostic performance of pathology laboratories. The CIQC program has a high-quality design, due to the reasonably large number of samples in each PTA, using same test samples for a large number of laboratories and across ER and PR tests, and the repeated participation of laboratories. Our statistical modeling uses these key design features to maximize monitoring potential, by linking data from the same patient, laboratories, time points, and tests. Our statistical modeling also increases precision of accuracy estimates by using a single indicator value for overstaining or understaining by IHC, so individual assessment rounds from CIQC can be robustly interpreted. Monitoring laboratories over time can enhance feedback to facilitate improvement in performance, while minimizing overcalling or undercalling laboratory performance.

REFERENCES

1. Hammond ME, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Arch Pathol Lab Med*. 2010;134:e48–e72.
2. Barnes DM, Millis RR, Beex LV, et al. Increased use of immunohistochemistry for oestrogen receptor measurement in mammary carcinoma: the need for quality assurance. *Eur J Cancer*. 1998;34:1677–1682.
3. Rhodes A, Jasani B, Balaton AJ, et al. Immunohistochemical demonstration of oestrogen and progesterone receptors: correlation of standards achieved on in house tumours with that achieved on external quality assessment material in over 150 laboratories from 26 countries. *J Clin Pathol*. 2000;53:292–301.
4. Torlakovic EE, Riddell R, Banerjee D, et al. Canadian Association of Pathologists-Association canadienne des pathologistes National Standards Committee/Immunohistochemistry: best practice recommendations for standardization of immunohistochemistry tests. *Am J Clin Pathol*. 2010;133:354–365.
5. Lester SC, Bose S, Chen YY, et al. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch Pathol Lab Med*. 2009;133:1515–1538.
6. Makretsov N, Gilks CB, Alaghehbandan R, et al. Development of an evidence-based approach to external quality assurance for breast cancer hormone receptor immunohistochemistry: comparison of reference values. *Arch Pathol Lab Med*. 2011;135:874–881.
7. Terry J, Torlakovic EE, Garratt J, et al. Implementation of a Canadian external quality assurance program for breast cancer biomarkers: an initiative of Canadian Quality Control in immunohistochemistry (cIQc) and Canadian Association of Pathologists (CAP) National Standards Committee/Immunohistochemistry. *Appl Immunohistochem Mol Morphol*. 2009;17:375–382.
8. Wolff AC, Hammond ME, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *Arch Pathol Lab Med*. 2014;138:241–256.
9. Phillips T, Murray G, Wakamiya K, et al. Development of standard estrogen and progesterone receptor immunohistochemical assays for selection of patients for antihormonal therapy. *Appl Immunohistochem Mol Morphol*. 2007;15:325–331.
10. Nofech-Mozes S, Vella ET, Dhesy-Thind S, et al. Systematic review on hormone receptor testing in breast cancer. *Appl Immunohistochem Mol Morphol*. 2012;20:214–263.
11. Wludarski SC, Lopes LF, Duarte IX, et al. Estrogen and progesterone receptor testing in breast carcinoma: concordance of results between local and reference laboratories in Brazil. *Sao Paulo Med J*. 2011;129:236–242.
12. Ochodo EA, Reitsma JB, Bossuyt PM, et al. Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. *J Clin Epidemiol*. 2013;66:1281–1288.
13. Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata J*. 2009;9:211–229.
14. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol*. 2006;59:1331–1332. Author reply 32-3.
15. Rabe-Hesketh S, Skrondal A. *Multilevel and Longitudinal Modeling Using Stata*, 2nd ed. College Station, TX: Stata Press; 2008.
16. Liang KY, ZEGER SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
17. Gibbons LE, Hosmer DW. Conditional logistic regression with missing data. *Commun Stat Simul Comput*. 1991;20:109–120.
18. Carey VJ, Lumley T, Ripley B. *Generalized Estimation Equation Solver*. 2012.Available at: http://CRAN.R-project.org/package = gee. R package version 4.13-18