Poorly Performing Physicians: Does the Script Concordance Test Detect Bad Clinical Reasoning?

FRANÇOIS GOULET, MD, MA; ANDRÉ JACQUES, MD; ROBERT GAGNON, MPSY;
BERNARD CHARLIN, MD, PHD; ABDO SHABAH, MD

Introduction: *Evaluation of poorly performing physicians is a worldwide concern for licensing bodies. The Collège des Médecins du Québec currently assesses the clinical competence of physicians previously identified with potential clinical competence difficulties through a day-long procedure called the Structured Oral Interview (SOI). Two peer physicians produce a qualitative report. In view of remediation activities and the potential for legal consequences, more information on the clinical reasoning process (CRP) and quantitative data on the quality of that process is needed. This study examines the Script Concordance Test (SCT), a tool that provides a standardized and objective measure of a specific dimension of CRP, clinical data interpretation (CDI), to determine whether it could be useful in that endeavor.*

Methods: *Over a 2-year period, 20 family physicians took, in addition to the SOI, a 1-hour paper-and-pencil SCT. Three evaluators, blind as to the purpose of the experiment, retrospectively reviewed SOI reports and were asked to estimate clinical reasoning quality. Subjects were classified into 2 groups (below and above median of the score distribution) for the 2 assessment methods. Agreement between classifications is estimated with the use of the Kappa coefficient.*

Results: *Intraclass correlation for SOI was 0.89. Cronbach alpha coefficient for the SCT was 0.90. Agreement between methods was found for 13 participants (Kappa: 0.30, P = 0.18), but 7 out of 20 participants were classified differently in both methods. All participants but 1 had SCT scores below 2 SD of panel mean, thus indicating serious deficiencies in CDI.*

Discussion: *The finding that the majority of the referred group did so poorly on CDI tasks has great interest for assessment as well as for remediation. In remediation of prescribing skills, adding SCT to SOI is useful for assessment of cognitive reasoning in poorly performing physicians. The structured oral interview should be improved with more precise reporting by those who assess the clinical reasoning process of examinees, and caution is recommended in interpreting SCT scores; they reflect only a part of the reasoning process.*

Key Words: *clinical reasoning, clinical competence, assessment, Script Concordance Test, poorly performing physicians*

Introduction

Assessment of physician competence is a worldwide concern.[1,2] Many medical and specialty boards and regulatory authorities have developed specific programs to assess the performance of practicing physicians.[2–9] Many of these programs share the same assessment tool, the peer-review process, which is well recognized for its face validity, but also known for some questionable reliability issues.[10–15] Other programs have developed specific assessment tools, such as the physician achievement review or the multi-source assessment in Alberta and the Maritime Provinces in Canada, the PREP in Ontario, or by direct or video observation in Netherlands.[4,16–20] Finucane et al.[5] defined 3 levels of performance assessment: Level 1—screening of the whole population of physicians, Level 2—screening for difficulties among specific groups of physicians considered at higher risk of poor performance, and Level 3—targeted diagnostic intervention for physicians for whom great concerns of performance have been expressed.[5] In the United States, some organizations, such as Albany Medical College, University of Wisconsin, the Center for Personalized Education for Physicians (CPEP) in Denver, and the

Disclosure: All authors report that the conduct of the presently reported study was supported by the Medical Council of Canada.

Dr. Goulet: Assistant Director, Practice Enhancement Division, Collège des médecins du Québec; *Dr. Jacques:* Director, Practice Enhancement Division, Collège des médecins du Québec; *Mr. Gagnon:* Manager, Evaluation Bureau, CPASS, Faculty of Medicine, Université de Montréal; *Dr. Charlin:* Director, Research & Development, CPASS, Faculty of Medicine, Université de Montréal; *Dr. Shabah:* 8867 Avenue de Chateaubriand, Montreal.

Correspondence: François Goulet, Practice Enhancement Division, Collège des médecins du Québec, 2170, René-Lévesque Ouest, Montréal, Québec H3H 2T8, Canada; e-mail: goulet.cmq@sympatico.ca.

• Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/chp.20076

Physician Assessment and Clinical Education (PACE) at the University of California, provide Level 3 assessment and/or remediation.[21]

The Collège des médecins du Québec (CMQ) is the regulatory authority that issues licenses to practice medicine in the province of Quebec. It also oversees the quality of medical practice and is responsible for ensuring that physicians remain competent. For physicians identified as underperforming, or wanting to return to practice after a long interruption, the CMQ uses a Level 3 assessment procedure called the Structured Oral Interview (SOI). The SOI is comprised of a series of simulated medical encounters based on written material. Two peer physicians evaluate clinical data gathering, clinical procedures, interpretation of complementary tests, pharmacological treatment, and patient follow-up. Assessment of competence is highly structured and based on the identification of cases' key features by the physician under evaluation. At the end of SOI, assessors provide a written report that documents physicians' areas of strength and weakness. Quality and usefulness of the SOI assessment in this context are documented, as there is satisfying evidence of face validity, and as content validity is ensured by the use of well-designed table of specification. Reliability is well documented, and agreement between evaluators is high (91%).[22,23] Nevertheless, in view of developing remediation activities as well as the potential for legal consequences, more information on the clinical reasoning process (CRP) and quantitative data on the quality of the process are needed.

The Script Concordance Test (SCT) is a relatively new tool designed to measure a specific but crucial element of clinical reasoning: clinical data interpretation (CDI).[24] It has shown evidence of validity and reliability, and is based on a cognitive psychology theory of clinical competence.[25,26] It puts examinees in authentic written clinical situations in which they must judge the effect new data has on the status of specified options.[27] One significant characteristic of the SCT format is that it allows testing in ill-defined contexts that are often typical of practice. Calculation of scores on the test reflects the degree of concordance existing between examinees' answers and those of a panel of reference.[28] Studies in gynecology, radiology, family medicine, and surgery have shown a high degree of reliability and support for construct validity, with lowest mean scores for medical students, intermediate scores for residents, and higher scores for faculty.[25,28–30]

Hauer recently reported that there are very few studies on the detection and remediation of doctors with clinical reasoning difficulties.[31] In this perspective, the purpose of the study was to determine if a 1-hour SCT would complement SOI by providing more information on the clinical reasoning process (CRP) and quantitative data on the quality of that process. Positive findings may have implications for assessment strategies of poorly performing physicians.

Methods

Participants

Over a 2-year span, 25 physicians underwent the Level 3 assessment (SOI) in family medicine. Most of them were referred for suspicion of potential performance problems, but others were physicians who were looking to resume their practice after a long period of clinical inactivity, physicians who planned to change the scope of their practice (from surgery to family medicine, for example), or physicians who thought they may have weaknesses or deficiencies (voluntary assessment). Accordingly, competence among participants ranged widely. The study was approved by an ethics board. The addition of an SCT to the usual procedure was explained to participants, who were then asked to participate in the study. Two physicians refused to participate, and 3 others did not complete the SCT. As a result, the study included 20 participants.

Instruments

In any clinical case, there are a few unique, essential elements in decision making that, alone or in combination, are the critical steps in the successful resolution of the clinical problem (the Key Features).[32,33] Both SOI and SCT questions were built on these premises, and both were specifically designed to assess family physicians.

SOI. To carry out tasks involved in clinical competence evaluation, the CMQ uses a pool of specially trained family physician assessors (FPAs). SOI is a standardized and highly structured test.[22] It consists of a day-long encounter between an examinee and 2 FPAs. Examinees work orally through 20 clinical cases selected from a pool of 42. A few cases use simulated patients to check history taking, physical examination, and patient–doctor relationship specifically. There is no time constraint and cues are provided at each step of the process to be sure that examinees are not headed in a wrong direction. At the end of the interview, the 2 FPAs produce a written report of approximately 12–15 pages, called the performance profile. This report is designed to guide remediation, and it includes a subjective evaluation of the quality of clinical reasoning (data collection, generation of appropriate differential diagnosis, action justification, having an overall sense of the problem).

SOI Clinical Reasoning Scores. Three family physicians working at the CMQ (evaluators) carried out retrospective analysis of SOI reports. Their task was to review the reports, extract information concerning clinical reasoning and grade its level on a 3-point scale (acceptable, cause for concern, unacceptable). *Acceptable* (score = 3) meant that evidence demonstrated the doctor's performance was above the standard for fitness to practice. *Cause for concern* (score = 2) meant there were concerns but there was not sufficient

DOI: 10.1002/chp

evidence to suggest seriously deficient performance. *Unacceptable* (score = 1) indicated there was evidence of repeated failure to comply with the professional standards. This simple 3-point scale was deemed sufficiently precise and easy to use by the 3 assessors. Assessments were made blinded. For each examinee, SOI clinical reasoning score (SOI-CR) represents the mean value given by the 3 evaluators.

SCT. SCT was based on a bank of items with documented evidence of reliability and validity in family medicine.[28] The administered test consisted of 16 vignettes/145 items. Half of the SCT was administered before each half-day SOI session. Each half was completed in approximately 30 minutes. To prevent sequence effect, the order of presentation of each half SCT was randomly determined. The scoring key was developed with the use of the answers of a convenience sample (the reference panel) of 13 family physicians in office practice without any university affiliation. The sample comprised the 5 members of the FPAs pool and 8 physicians recruited through a snowball strategy.[28]

SCT Optimization. From the administered SCT, 3 strategies were used to retain only the best items: (a) calculation of effect size of each item (difference of mean score of physician and panel members divided by panel standard deviation) to eliminate items with effect size smaller than 0.25; (b) items with variance in the panel higher than 1 were excluded; and (c) item–total correlation was used to retain only items with positive correlation.

Statistical Analyses

Reliability. Reliability of clinical reasoning SOI scores among the 3 evaluators was estimated with intraclass correlation, whereas reliability of SCT was estimated with Cronbach alpha coefficient of internal coherence.

Tool Comparison. The average value of the 3 raters on the quality-of-reasoning scale was used to score the reasoning aspect in SOI. Concordance between SOI-CR and SCT scores was analyzed with Kappa, after dichotomization of scores with the use of the median value of each distribution of scores as the cutoff point. The small sample size precluded the utilization of finer categorization. Ninety-five percent confidence intervals are calculated for Kappa and ICC. All statistics are 2-tailed.

Results

Comparison of assessment averages of clinical reasoning on SOI reports by the 3 raters led to an ICC of 0.89. FIGURE 1 shows the distribution of participants' SOI clinical reasoning scores. Five participants (25%) had a perfect score of 3. Mean value of the distribution is 2.13 (SD = 0.72) and the median score is 2.17.

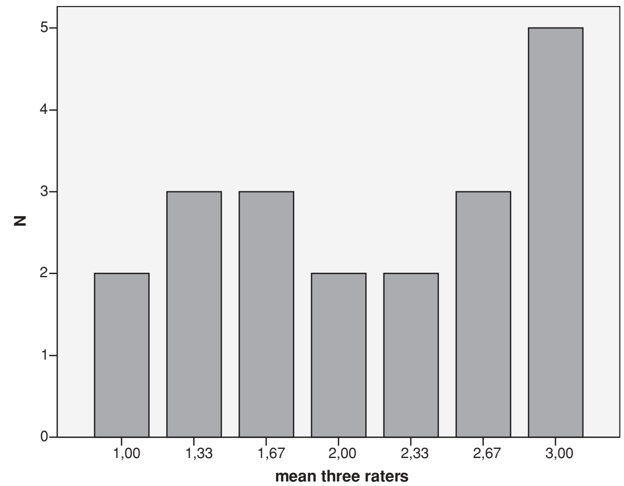

FIGURE 1. Mean rating of the 3 clinical reasoning raters from the SOI reports (SOI-CR scores)

After optimization, SCT had 13 vignettes/64 items. Items covered data interpretation on history taking (19 items), physical exam (26 items), diagnosis and interpretation of tests (7 items), and treatment (12 items). Value of the Cronbach alpha coefficient was 0.90. Mean and standard deviation of SCT scores for the group of 20 participants were 52.3 and 15.8. To make SCT scores meaningful, participants' scores need to be compared with panel's mean (82.1) and standard deviation (4.1). Panel's standard deviation (SD) can be considered a yardstick that depicts how far participants are from panel mean. Within this group of participants referred for potential clinical competence problems or seeking to return to practice after an interruption, none had a score above panel mean: 1 was between panel mean and −1 SD, 3 were between −2 and 3 SD, 1 was between −3 and 4 SD, and 15 were below 5 SD of panel mean.

FIGURE 2 illustrates the scattergram of scores on both measures. SCT mean scores (and SD) were 43.7 (12.8) for the group having SOI-CR scores under median and 59.3 (15.0) for the group having SOI-CR scores over median. Associated ICC value is 0.26 (CI = −0.20 to 0.62). There was an agreement between tools for 13 participants (65%). Calculated Kappa value is 0.30 ($P = 0.18$; CI = −0.12 to 0.72). Among the 7 discordant cases, 3 participants had SCT scores more than 5 SD below panel mean (44.9, 43.0, and 36.2), whereas they scored above median (2.17) on the SOI-CR. Four participants who scored low on SOI-CR (1.0, 1.3, 1.7, 1.7) scored above median on SCT (scores = 54.5, 57.8, 57.4, and 63.7).

Discussion

Three physicians who are employed at the College with responsibility to ensure that physicians maintain competency were asked, retrospectively, to review SOI reports, to make judgments about the quality of clinical reasoning, and to determine if performance was sufficient to classify the doctor as

DOI: 10.1002/chp

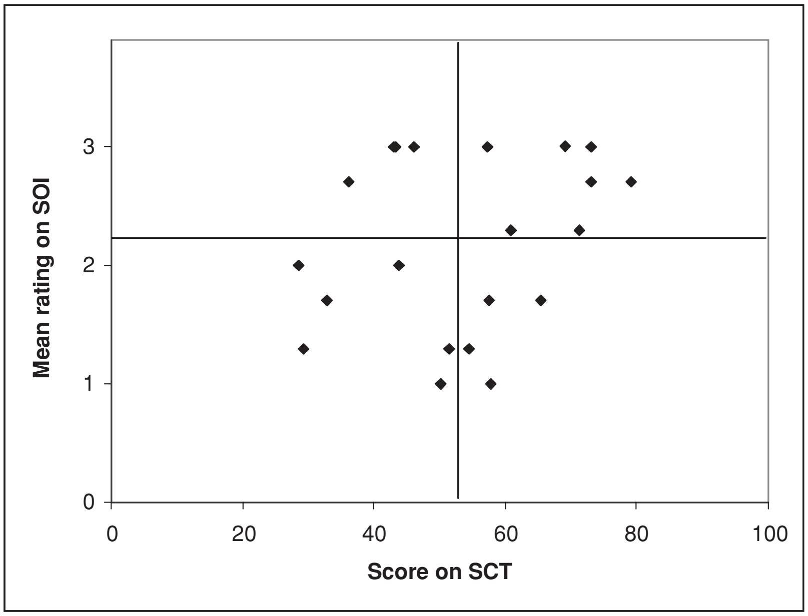

FIGURE 2. Mean rating scattergram of 3 raters on SOI and global rating on SCT. Cross lines show median values.

competent for practice. The average value of these judgments showed good interjudge reliability with ICC value of 0.89. The SCT specifically designed for use with poorly performing physicians show high reliability for an hour of testing time. All participants but 1 had SCT scores below 2 SD of panel mean, thus indicating deficiencies in CDI, while median SOI-CR score was 2.17, that is, above the mark "cause for concern."

By design, SOI measures multiple tasks of the clinical encounter, and SCT measures a specific aspect of the clinical reasoning process: data interpretation. Context is given; data and hypotheses are provided. The task for participants is to decide the effect that new data have on the status of the given hypotheses. This very focused clinical task requires organized knowledge and provides indications on clinical judgment quality.[34] Given these design differences, it is not really surprising that the results were neither consistent nor consistently different.

The study does highlight the multifaceted complexity of the clinical reasoning process. As the SCT has been used in many studies and has shown a fair degree of validity in classifying trainees according to their level of expertise and clinical reasoning,[25–29,34] results of the present study suggest that, for the goal of remediation prescription, adding an SCT to the SOI is useful for the assessment of clinical reasoning in poorly performing physicians. Though study results also underscore that care should be taken in interpretation of SCT scores. They reflect only a part of the whole reasoning process. Results also indicate that SOI procedure should be improved, with request for assessors to document and report more precisely their observations on the clinical reasoning process of the examinees.

According to the Kappa analysis, 7 out of 20 participants were classified differently in both methods. It is interesting to examine these disagreements. In 4 cases, SCT scores indicated above-median performance on CDI, whereas SOI-CR scores were below median, thus indicating that clinical reasoning difficulties may lay in other elements, such as strategies for data collection, hypotheses generation, or closure of the reasoning process. In the 3 other cases, SOI-CR scores were above median, whereas SCT scores indicated a potential problem with CDI. Beside this, although being unable to document this formally, we strongly suspect that the 5 physicians who refused participation or did not complete the SCT perceived SCT cognitive tasks as too difficult or too unusual for them, thus indicating potential serious CDI difficulties.

Some important limitations in the present study are to be noted. Although the SOI procedure is highly standardized, the report written by the 2 FPAs is not: There is no clear instruction to document the reasoning aspects of the physician under observation. This variability in reporting reasoning may have led to underreporting problems. Furthermore, retrospective assessment of clinical reasoning quality was not an easy task, and we were limited to using a simple qualitative 3-point scale. After many attempts to transform this scale into something meaningful and statistically useful, we chose to report scores as below or over the median, while being conscious of the limits of this modification. Sample size is clearly nonoptimal for estimation of concordance; lack of power means difficulty showing true concordance between both methods. It must be said that this study was conducted in the context of day-to-day data gathering, and that the CMQ uses the SOI procedure about 8 to 12 times a year.

DOI: 10.1002/chp

Conclusion

The use of an objective and standardized measure of CDI in the assessment process of physicians with performance problems is promising. Data suggest that the SCT measures some aspects of the clinical reasoning process of physicians under evaluation that are not commonly detected by peer assessment in the context of an SOI. The finding that the majority of the referred group did so poorly relative to the reference group on the SCT is of great interest both for assessment as well as remediation.

Lessons for Practice

- The structured oral interview measures multiple tasks of the physician's clinical practice; the script concordance test measures a specific aspect of the clinical reasoning process, data interpretation.
- For underperforming physicians, the script concordance test measures aspects of the clinical reasoning process that are not usually detected in a structured oral interview.
- The structured oral interview should be improved with more precise reporting by those who assess the clinical reasoning process of examinees.

References

1. Davies HT, Shields AV. Public trust and accountability for clinical performance: lessons from the national press reportage of the Bristol hearing. *J Eval Clin Pract*. 1999;5:335–342.
2. Southgate L, Campbell M, Cox J, Foulkes J, Jolly B, McCrorie P, Tombleson P. The General Medical Council's Performance Procedures: the development and implementation of tests of competence with examples from general practice. *Med Educ*. 2001;35(S1):20–28.
3. Eliasson G, Berg L, Carlsson P, Lindstrom K, Bengtsson C. Facilitating quality improvement in primary health care by practice visiting. *Qual Health Care*. 1998;7:48–54.
4. Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. *JAMA*. 1993;270:1046–1051.
5. Finucane PM, Bourgeois-Law GA, Ineson SL, Kaigas TM. A comparison of performance assessment programs for medical practitioners in Canada, Australia, New Zealand, and the United Kingdom. *Acad Med.* 2003;78:837–843.
6. Norton PG, Dunn EV, Soberman L. What factors affect quality of care? Using the Peer Assessment Program in Ontario family practices. *Can Fam Physician*. 1997;43:1739–1744.
7. Southgate L, Cox J, David T, et al. The General Medical Council's Performance Procedures: peer review of performance in the workplace. *Med Educ*. 2001;35(S1):9–19.
8. St George K, Kaigas T, McAvoy P. Assessing the competence of practicing physicians in New Zealand, Canada, and the United Kingdom: progress and problems. *Fam Med*. 2004;36:172–177.
9. Goulet F, Jacques A, Gagnon R, et al. Performance assessment. Family physicians in Montreal meet the mark! *Can Fam Physician*. 2002;48:1337–1344.
10. Wu L, Ashton CM. Chart review. A need for reappraisal. *Eval Health Prof*. 1997;20:146–163.
11. Lockyer J, Harrison V. Performance assessment: the role of chart review analysis and CME. In: Davis DA, Fox RD, Eds. *The Physician as Learner: Linking Research to Practice*. Chicago, IL: American Medical Association; 1994:169–186.
12. Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical performance? A study using simulated patients. *Br J Gen Pract*. 1994;44:153–156.
13. Tugweel P, Dok C. Medical record review. In: Neufeld VR, Norman GR, Eds. *Assessing Clinical Competence*. New York, NY: Springer-Verlag; 1985:142–182.
14. Goldman RL. The reliability of peer assessments: a meta-analysis. *Eval Health Prof*. 1994;17:3–21.
15. Smith MA, Atherly AJ, Kane RL, Pacala JT. Peer review of the quality of care. Reliability and sources of variability for outcome and process assessments. *JAMA*. 1997;278:1573–1578.
16. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. *CMAJ*. 1999;161:52–57.
17. Lockyer J. Multisource feedback in the assessment of physicians competencies. *J Contin Educ Health Prof*. 2003;23:4–12.
18. College of Physicians and Surgeons of Alberta. College programs. *Physician Achievement Review (PAR) program: PAR evaluation*. www.par-program.org/PAR-Info.htm. Accessed August 2010.
19. Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and coworker raters to a multisource feedback process: a pilot study. *Acad Med*. 2003;78(Suppl):S42–S44.
20. Ram P, van der Vleuten C, Rethans JJ, Grol R, Aretz K. Assessment of practicing family physicians: comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. *Acad Med*. 1999;74:62–69.
21. Humphrey C. Assessment and remediation for physicians with suspected performance problems: an international survey. *J Contin Educ Health Prof*. 2010;30(1):26–36.
22. Jacques A, Sindon A, Bourque A, Bordage G, Ferland JJ. Structured oral interview. One way to identify family physicians educational needs. *Can Fam Physician*. 1995;41:1346–1352.
23. Miller F, Jacques A, Brailovsky C, Sindon A, Bordage G. When to recommend compulsory versus optional CME programs? A study to establish criteria. *Acad Med*. 1997;72:760–764.
24. Reed GW, Debra LK, Hoffman R. Medical student acquisition of clinical working knowledge. *Teach Learn Med*. 2008;20(1):5–10.
25. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof*. 2004;27:304–319.
26. Charlin B, Boshuizen HPA, Custers EJFM, Feltovich PJ. Scripts and clinical reasoning. *Med Educ*. 2007;41:1178–1184.
27. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: a tool to assess the reflective clinician. *Teach Learn Med*. 2000;12:189–195.
28. Charlin B, Gagnon R, Pelletier J, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ*. 2006;40:848–854.
29. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ*. 2005;39:284–291.
30. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. Stability of clinical reasoning assessment results with the Script Concordance Test across two different linguistic, cultural and learning environments. *Med Teach*. 2002;24:522–527.

DOI: 10.1002/chp

31. Hauer K, Ciccone A, Henzel TR, et al. Remediation of the deficiencies of physicians across the continuum from medical school to practice: a thematic review of the literature. *Acad Med*. 2009;84:1822–1832.
32. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med*. 1995;70:194–201.
33. Farmer EA, Page G. A practical guide to assessing clinical decision making skills using the key features approaches. *Med Educ.* 2005;39:1188–1194.
34. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teach Learn Med.* 2010;22:180–186.

DOI: 10.1002/chp