Title

Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios?

Journal

International Journal of Emergency Medicine, 7(1)

ISSN

1865-1372

Authors

Ahmadi, Seyed-Foad
Khoshkish, Shahin
Soltani-Arabshahi, Kamran
et al.

Publication Date

2014-12-01

DOI

10.1186/s12245-014-0034-3

International Journal of Emergency Medicine
a SpringerOpen Journal

BRIEF RESEARCH REPORT Open Access

Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios?

Seyed-Foad Ahmadi[1,2], Shahin Khoshkish[1,3], Kamran Soltani-Arabshahi[1], Peyman Hafezi-Moghadam[4], Golara Zahmatkesh[1], Parisa Heidari[1,5], Davood Baba-Beigloo[6], Hamid R Baradaran[1] and Shahram Lotfipour[7*]

Abstract

Background: We aimed to compare the clinical judgments of a reference panel of emergency medicine academic physicians against evidence-based likelihood ratios (LRs) regarding the diagnostic value of selected clinical and paraclinical findings in the context of a script concordance test (SCT).

Findings: A SCT with six scenarios and five questions per scenario was developed. Subsequently, 15 emergency medicine attending physicians (reference panel) took the test and their judgments regarding the diagnostic value of those findings for given diseases were recorded. The LRs of the same findings for the same diseases were extracted from a series of published systematic reviews. Then, the reference panel judgments were compared to evidence-based LRs. To investigate the test-retest reliability, five participants took the test one month later, and the correlation of their first and second judgments were quantified using Spearman rank-order coefficient.
In 22 out of 30 (73.3%) findings, the expert judgments were significantly different from the LRs. The differences included overestimation (30%), underestimation (30%), and judging the diagnostic value in an opposite direction (13.3%). Moreover, the score of a hypothetical test-taker was calculated to be 21.73 out of 30 if his/her answers were based on evidence-based LRs.
The test showed an acceptable test-retest reliability coefficient (Spearman coefficient: 0.83).

Conclusions: Although SCT is an interesting test to evaluate clinical decision-making in emergency medicine, our results raise concerns regarding whether the judgments of an expert panel are sufficiently valid as the reference standard for this test.

Keywords: Clinical judgment; Script concordance test; Likelihood ratio; Visual analog scales; Evidence-based medicine; Diagnosis; Decision-making

Introduction

Script concordance test (SCT) has become a recognized tool to assess clinical reasoning in various fields including emergency medicine [1-14]. This case-based test consists of short clinical scenarios followed by questions regarding diagnosis or management. The questions are presented in three parts: A) a diagnostic or management option, B) a clinical finding, and C) a scale to capture examinee's decision (Figure 1) [15]. The test is based on measuring the concordance of test-taker judgments with those of a reference panel of experts [15]. Expert physicians usually organize their knowledge regarding diseases in 'illness scripts', and when they encounter patients, they effortlessly recall the relevant scripts and promptly recognize the most appropriate courses of action [16]. SCT is indeed an effort to capture how close the scripts of test-takers are with the scripts of experts, and the rationale behind it is the more close to the

* Correspondence: shl@uci.edu
[7]Department of Emergency Medicine, School of Medicine, University of California Irvine, Irvine, CA 92697, USA
Full list of author information is available at the end of the article

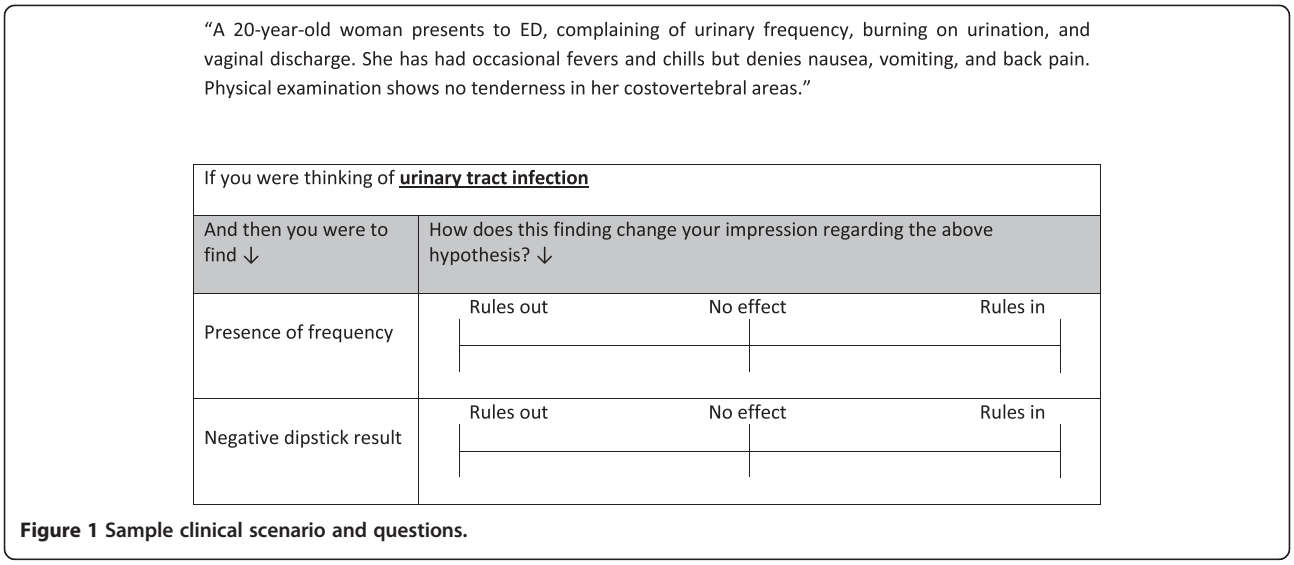

Figure 1 Sample clinical scenario and questions.

experts' scripts, the better the decision-making by the test-takers. However, the expert judgments are reported to be frequently incorrect [17] and therefore, the reference standard of the test, which is the expert judgments, seems to be not necessarily valid. The test is mainly used to assess reasoning in uncertain situations [15] in which robust evidence is usually limited. However, that the test reference standard is not necessarily valid is still a critical issue and should be carefully investigated.

The diagnostic value of clinical and/or paraclinical findings is an appropriate context in which expert opinions can be compared with the best current evidence. At one hand, findings can be presented to experts and how such findings would modify the experts' diagnostic judgments, regarding the likelihood of particular diseases, can be measured. On the other hand, the expected effect of the presented findings on the likelihood of the same disease can be sought from the best current evidence. According to Bayes' theorem, the likelihood ratio (LR) of any diagnostic finding is a precise indicator of the expected change in the likelihood of that disease if the suspected individual has that particular finding [18]. Fortunately, LRs for a wide variety of clinical and paraclinical findings are either available or calculable based on robust studies [19]. Hence, we aimed to seek the judgments of a panel of emergency medicine experts regarding the diagnostic value of select clinical and paraclinical findings, acquire the evidence-based LRs for the same findings, and finally compare the judgments against the LRs.

Methods

Study design and setting

We invited all emergency medicine attending physicians (consultants) of the main teaching hospitals of two academic universities (Iran University of Medical Sciences and Tehran University of Medical Sciences) to participate in our study. The two teaching hospitals have an ED yearly census of over 90,000 patients. Participating consultants consented to be enrolled after receiving detailed explanations regarding the purpose and the design of the study. The required sample size was 15 according to the SCT development guidelines [15].

Test development

We developed a test containing six clinical scenarios on the following emergent conditions: 1) meningitis, 2) myocardial infarction, 3) pneumonia (in a child), 4) thoracic aortic dissection, 5) appendicitis, and 6) congestive heart failure. Each scenario was followed by five questions, and each question was intended to measure the judgments of our panel of consultants regarding the diagnostic value of the presence or absence of a clinical, laboratory or imaging finding. To develop the test, two investigators (SK and SFA) studied a series of systematic reviews containing a wide variety of clinical scenarios and the corresponding evidence-based LRs for the related clinical and paraclinical findings [19]. The investigators selected the scenarios and findings that could properly represent diagnostic challenges in the emergency room. Afterwards, they designed a SCT based on those scenarios and findings according to the recommended guidelines [15]. The only variation from the ordinary SCT was using 10-cm visual analog scales (VASs) instead of five-point Likert scales since both tools yield comparable measurements [20,21] while VAS was also able to quantify the judgments of the reference panel. In addition to the main test, a separate sample scenario with two questions was developed and utilized to familiarize the participants with the test-taking process,

so that they could completely understand how the test works before taking the main test (Figure 1).

Test validation

Prior to the main experiment, the content validity of the test was carefully evaluated and confirmed by an emergency medicine expert (PHM) and a medical education expert (KSA). In addition, we invited five expert participants to take the test again one month after the main experiment in order to measure the test-retest reliability. The correlation of the two sets of responses was measured using Spearman rank-order coefficient.

Data collection

After a brief orientation, the consultants received the main test in their private office and answered the questions while having no access to any medical resources. To answer each question regarding the diagnostic value of a finding, they marked a point on the VAS. The numbers equivalent to the VAS markings were identified using an ordinary ruler.

Analytical approach

The numbers representing the consultants' judgments were multiplied by 2 in order to rescale the original range of −5 to 5 to a range of −10 to +10. For the LR values, LRs >10 or <0.1 were considered 10 and 0.1, respectively, as we needed to establish a bounded LR range. This conversion seemed rational as an LR = 10 is considered sufficiently large to rule-in a disease and an LR = 0.1 is considered sufficiently small to rule-out a disease [22], and whether an LR is 10 or higher, or whether it is 0.1 or lower is not substantially different for clinical reasoning purposes. Subsequently, LR values were converted to '$10 \times log(LR)$' in order to convert their naturally geometric scale to an arithmetic scale. We used one-sample t test to compare the transformed mean judgments with the corresponding transformed LRs. In addition, the score of a hypothetical test-taker was calculated if his/her answers were based on evidence-based LRs, and the answers were scored based on the judgments of our consultants as the reference standard. The calculation is described elsewhere [15]. For statistical analysis, IBM SPSS Statistics 19 was used. A $P < 0.05$ was considered significant.

Findings

Participant characteristics

Fifteen emergency medicine consultants consented to participate in our study, from which 13 consultants were board certified in emergency medicine and the other two consultants were board certified in internal medicine and pediatrics, respectively, with additional fellowship training in emergency medicine. The mean age, clinical practice experience, and emergency medicine experience were 35.9, 10.3, and 6.6 years, respectively. The Spearman coefficient was 0.83 for the two sets of answers from a subset of five consultants.

Comparison of the reference panel judgments against the evidence-based LRs

We have summarized the results of comparing values representing consultants' judgments with evidence-based LRs in Table 1. Our results showed that in 22 out of 30 (73.3%) findings, the mean judgments were significantly different from the corresponding LRs. Our results also demonstrated that consultants overestimated the value of the 9 (30%) findings and underestimated the value of another 9 (30%) findings. In addition to the above discrepancies regarding the magnitude of the diagnostic value, the consultants chose a different direction (regarding ruling in or ruling out) for 4 (13.3%) findings compared to the evidence-based LRs.

Subgroups of positive and negative findings

When positive and negative findings (presence or absence of findings) were considered separately, we noted a significant difference between the consultants' judgments and the LRs in 17 out of 20 (85%) positive findings and 5 out of 10 (50%) negative findings. The diagnostic values of 8 (40%) positive and 1 (10%) negative findings were overestimated and the values of 7 (35%) positive and 2 (20%) negative findings were underestimated by the consultants. Moreover, the judgments were in opposite direction to the LRs in 2/20 (10%) and 2/10 (20%) of the positive and negative findings, respectively.

Subgroups of history, physical examination, and laboratory findings

When we calculated the percentage of significantly different consultants' judgments from the corresponding LRs in subgroups of findings from history, physical examination, and laboratory/imaging findings, we observed comparable percentages for findings of history (6 out of 8: 75.0%), physical examination (10 out of 14: 71.4%), and laboratory/imaging (6 out of 8: 75%). However, physical examination findings were more frequently overestimated (25%, 35.7%, and 25% for history, physical examination, and laboratory findings, respectively) and less frequently underestimated (37.5%, 21.4%, and 37.5% for history, physical examination, and laboratory findings, respectively).

The score of the hypothetical test-taker

The calculated score of a hypothetical test-taker was 21.73 out of 30 based on the consultants' judgments as the reference standard. The categorization of LRs, the

Table 1 Comparison of the participants' judgments with likelihood ratios (LRs)

Disease/finding	Judgments Mean (SD)	LR: raw [Transformed]	*P* value[a]	Difference
1. Meningitis				
1A. Presence of headache	*4.08* (2.23)	1.10 [*0.41*]	$P < 0.001$	*Overestimation* [+][b]
1B. Absence of nausea/vomiting	*0.16* (0.87)	0.64 [*−1.93*]	$P < 0.001$	*Contradictory*
1C. Presence of neck stiffness	*7.03* (2.17)	1.10 [*0.41*]	$P < 0.001$	*Overestimation* [+]
1D. Presence of Brudzinski's sign	*6.00* (2.72)	0.97 [*−0.13*]	$P < 0.001$	*Contradictory*
1E. Presence of Kernig's sign	*6.51* (1.69)	0.97 [*−0.13*]	$P < 0.001$	*Contradictory*
2. Myocardial infarction				
2A. Presence of chest pain with radiation to both arms	*4.59* (2.63)	4.10 [*6.12*]	$P = 0.043$	*Underestimation* [+]
2B. Absence of nausea/vomiting	*−0.21* (0.80)	0.87 [*−0.60*]	$P = 0.091$	-
2C. Presence of sharp chest pain or stabbing	*−0.47* (4.00)	0.30 [*−5.22*]	$P = 0.002$	*Underestimation* [–]
2D. Presence of any ST segment elevation	*7.62* (2.52)	3.20 [*5.05*]	$P = 0.002$	*Overestimation* [+]
2E. Presence of any Q wave	*4.21* (3.40)	3.90 [*5.91*]	$P = 0.074$	-
3. Pneumonia (*in a child*)				
3A. Presence of retraction	*5.31* (2.88)	1.00 [*0.00*]	$P < 0.001$	*Overestimation* [+]
3B. Absence of tachypnea	*−3.06* (4.70)	0.97 [*−0.13*]	$P = 0.030$	*Overestimation* [–]
3C. Presence of crackles	*4.96* (2.46)	1.60 [*2.04*]	$P = 0.001$	*Overestimation* [+]
3D. Presence of grunting	*4.00* (2.74)	2.70 [*4.31*]	$P = 0.665$	-
3E. Absence of fever	*−1.10* (3.02)	0.07 [*−1.19*]	$P = 0.920$	-
4. Thoracic aortic dissection				
4A. Presence of history of hypertension	*4.33* (2.27)	1.60 [*2.04*]	$P = 0.002$	*Overestimation* [+]
4B. Presence of focal neurologic deficit	*4.56* (2.60)	14.75 [10^c]	$P < 0.001$	*Underestimation* [+]
4C. Absence of pulse deficit	*0.45* (4.46)	0.70 [*−1.50*]	$P = 0.104$	-
4D. Absence of enlarged aorta/wide mediastinum in chest X-ray (CXR)	*2.01* (4.81)	0.30 [*−5.22*]	$P < 0.001$	*Contradictory*
4E. Absence of sudden chest pain	*−2.49* (4.19)	0.30 [*−5.22*]	$P = 0.025$	*Underestimation* [–]
5. Appendicitis				
5A. Absence of anorexia	*−2.81* (2.58)	0.64 [*−1.93*]	$P = 0.229$	-
5B. Presence of guarding	*4.90* (2.28)	1.70 [*2.30*]	$P = 0.001$	*Overestimation* [+]
5C. Absence of rebound tenderness	*−1.27* (3.91)	0.002 [-10^d]	$P < 0.001$	*Underestimation* [–]
5D. Presence of psoas sign	*3.80* (2.80)	2.40 [*3.80*]	$P = 0.999$	-
5E. White blood cell count of 12,000	*3.82* (3.14)	1.30 [*1.13*]	$P = 0.005$	*Overestimation* [+]
6. Congestive heart failure				
6A. Presence of third heart sound	*6.18* (2.15)	11.00 [10^e]	$P < 0.001$	*Underestimation* [+]
6B. Absence of cardiomegaly in CXR	*−3.89* (2.92)	0.33 [*−4.81*]	$P = 0.258$	-
6C. Presence of interstitial edema in CXR	*5.77* (2.17)	12.00 [10^f]	$P < 0.001$	*Underestimation* [+]
6D. Presence of atrial fibrillation in EKG	*2.73* (2.89)	3.80 [*5.79*]	$P = 0.001$	*Underestimation* [+]
6E. Presence of lateral EKG changes	*1.81* (2.15)	2.20 [*3.42*]	$P = 0.012$	*Underestimation* [+]

[a]Each *P* value is derived from a one-sample *t* test comparing the mean expert judgments (measured by visual analog scales) with the corresponding transformed likelihood ratio [$10 \times \log(LR)$].
[b]'Overestimation [+]' implies that the value of this finding was overestimated positively (i.e. towards ruling in).
[c]Since the highest possible value for the expert judgments was 10, the transformed LR of 11.69 was considered 10 in the corresponding one-sample *t* test.
[d]Similarly, the transformed LR of −25.32 was considered −10 in the corresponding one-sample *t* test.
[e]Likewise, the transformed LR of 10.41 was considered 10 in the corresponding one-sample *t* test.
[f]Also, the transformed LR of 10.79 was considered 10 in the corresponding one-sample t test.

Table 2 Calculation of the score of a hypothetical test-taker

Disease/finding		Score of the categories				
	Categorized LR	Very low (−2)	Low (−1)	Middle (0)	High (+1)	Very high (+2)
1. Meningitis						
1A. Presence of headache	*Middle (0)*	0	0	*0.57*	1	0.57
1B. Absence of nausea/vomiting	*Middle (0)*	0	0.07	*1*	0.07	0
1C. Presence of neck stiffness	*Middle (0)*	0	0.09	*0.09*	0.09	1
1D. Presence of Brudzinski's sign	*Middle (0)*	0	0	*0.22*	0.44	1
1E. Presence of Kernig's sign	*Middle (0)*	0	0	*0.12*	0.75	1
2. Myocardial infarction						
2A. Presence of chest pain with radiation to both arms	*Very high (+2)*	0	0	0.37	1	*0.50*
2B. Absence of nausea/vomiting	*Middle (0)*	0	0	*1*	0.07	0
2C. Presence of sharp chest pain or stabbing	*Low (−1)*	0.25	*1*	1	0.50	0.25
2D. Presence of any ST segment elevation	*High (+1)*	0.10	0	0.10	*0.30*	1
2E. Presence of any Q wave	*High (+1)*	0	0	1	*1*	1
3. Pneumonia (in a child)						
3A. Presence of retraction	*Middle (0)*	0	0	*0.28*	0.85	1
3B. Absence of tachypnea	*Middle (0)*	1	0.60	*0.80*	0.60	1
3C. Presence of crackles	*High (+1)*	0	0.16	0.33	*1*	1
3D. Presence of grunting	*High (+1)*	0	0	0.71	*1*	0.42
3E. Absence of fever	*Middle (0)*	0	0.71	*1*	0.28	0.14
4. Thoracic aortic dissection						
4A. Presence of history of hypertension	*High (+1)*	0	0.12	0.25	*1*	0.50
4B. Presence of focal neurologic deficit	*Very high (+2)*	0	0	0.25	1	*0.62*
4C. Absence of pulse deficit	*Middle (0)*	0.11	0.11	*1*	0.11	0.33
4D. Absence of enlarged aorta/wide mediastinum in CXR	*Low (−1)*	0	*1*	0.40	0.80	0.80
4E. Absence of sudden chest pain	*Low (−1)*	0.66	*0.33*	1	0.50	0
5. Appendicitis						
5A. Absence of anorexia	*Middle (0)*	0.25	0.50	*1*	0.12	0
5B. Presence of guarding	*High (+1)*	0	0.12	0.37	*1*	0.37
5C. Absence of rebound tenderness	*Very low (−2)*	*0.16*	1	0.66	0.33	0.33
5D. Presence of psoas sign	*High (+1)*	0	0	0.83	*1*	0.66
5E. WBC = 12,000	*Middle (0)*	0	0.12	*0.25*	1	0.50
6. Congestive heart failure						
6A. Presence of third heart sound	*Very high (+2)*	0	0.11	0	0.55	*1*
6B. Absence of cardiomegaly in CXR	*Low (−1)*	0.42	*1*	0.42	0.14	0.14
6C. Presence of interstitial edema in CXR	*Very high (+2)*	0	0	0.12	0.75	*1*
6D. Presence of atrial fibrillation in EKG	*High (+1)*	0	0.16	1	*0.83*	0.50
6E. Presence of lateral EKG changes	*High (+1)*	0	0	1	*0.66*	0
SUM				*21.73 (out of 30)*		

If a test-taker answered our script concordance test based on evidence-based likelihood ratios (LRs), and his/her answers were scored based on the judgments of this study's reference panel, the test-taker would get a score of 21.73 out of 30. For the above calculations, the numbers representing the judgments were categorized as very low, low, middle, high, and very high, similar to five-point Likert scales, using cutoff points of −6, −2, 2, and 6. Subsequently, the score of each category was calculated based on the judgments. Then, for LR of each finding, we identified its category ('categorized LR' column) and its corresponding score (italicized number). Finally, we added up all italicized numbers. The calculation of the scores based on the judgments of the reference panel is explained elsewhere [16].

score of each category, and the calculated score for each answer are summarized in Table 2.

Discussion

In summary, we observed that in a considerable proportion of the questions, the consultants' judgments regarding the value of the findings were significantly different from the corresponding evidence-based LRs; the differences included discrepant magnitude (over/underestimation) and also discrepant direction. Moreover, the value of the physical examination findings was more frequently overestimated and less frequently underestimated. This is possibly due to a popular attitude that the objective clinical findings are more valuable than paraclinical findings in the diagnosis process [23]. Furthermore, we showed that if a hypothetical test-taker had answered the test based on evidence-based LRs and his/her answers were evaluated using the consultants' judgments as a reference standard, the test-taker would get approximately two-thirds of the total score.

Previous studies have investigated aspects of SCT such as comparing the answer keys obtained from panels with different levels of expertise [24], optimizing the answer keys [25], improving the development of the scoring key [26], investigating the effect of variability within the reference panel [27], and validating the test in different clinical fields [1-14]. However, to our knowledge, no study had challenged the reference standard of SCT by evidence before our study. Clinical decision-making is a complex process influenced by both clinical knowledge and experience. As physicians collect experience by practicing medicine, their knowledge may be outdated [28,29]. Therefore, while the judgments of expert physicians benefit the most from valuable experiences, they may suffer from outdated knowledge and also cognitive biases [30-32]. A recent review has discussed the potential pitfalls of using SCT as a valid tool to measure clinical reasoning competencies, among which is implicitly discouraging the seeking of empirical evidence for the scoring key since this test assumes no single correct answer for any item [33].

Limitations

Despite the novel idea and methods, this study had the following limitations: A) Although the transformations in the judgment numbers and the LRs made these two entities comparable, the transformations could have introduced bias in the results. Knowing this limitation, we found no alternative approach to compare the consultant's judgments with evidence-based LRs. B) The optimal number of the clinical scenarios and the questions per each scenario is reported to be 20 and 3, respectively [15]. However, we used six scenarios and five questions per scenario because this test structure needed less time and could better address the time limitations of our consultant participants. C) The results were derived from only two centers in Tehran and therefore they cannot be easily generalized to all settings. D) As this study was carried out in emergency medicine context that has substantial differences with other specialties, our findings cannot be directly extrapolated to other fields of clinical medicine.

Conclusions

SCT is an interesting tool to score the clinical decision-making practices of novice trainees based on the judgments of expert physicians. However, experts' judgments may occasionally be inconsistent with evidence. This should raise concerns regarding the validity of the experts' judgments as a valid reference standard for SCT. We suggest future investigators should explore alternative evidence-based approaches to establish more robust reference standards for clinical reasoning tests such as the script concordance test in the field of emergency medicine.

Abbreviations

LR: likelihood ratio; SCT: script concordance test; VAS: visual analog scale..

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SFA conceived of the study, participated in the design of the study, performed the statistical analysis, and drafted the manuscript. SK conceived of the study, participated in the design of the study, and contributed to the data collection, statistical analysis, and draft of the manuscript. KSA contributed to conceive and design of the study and critically reviewed and revised the manuscript. PHM participated in design of the study and critically reviewed and revised the manuscript. GZ contributed to the data collection and draft of the manuscript. PH contributed to the data collection and data analysis. DBB contributed to the data collection and draft of the manuscript; HRB contributed to conceive of the study and critically reviewed and revised the manuscript. SL critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

Authors' information

SL is a professor of emergency medicine and public health at the University of California, Irvine's School of Medicine. PHM is an associate professor of emergency medicine and deputy for education at Iran University of Medical Sciences School of Medicine. KSA is a professor of medicine and medical education and head of the Department of Medical Education at Iran University of Medical Sciences. HRB is an associate professor of clinical epidemiology and evidence-based medicine at Iran University of Medical Sciences. The other authors were medical students at the time of conducting this study.

Acknowledgements

We would like to acknowledge Dr. Amir Nejati for his contributions in collecting data for this study.

Sources of funding

This study was the M.D. thesis of SK and was funded by Iran University of Medical Sciences. The authors have not received fund from any other source.

Author details

[1]Center for Educational Research in Medical Sciences, Iran University of Medical Sciences, Tehran 14496, Iran. [2]Program in Public Health, Department of Population Health and Disease Prevention, University of California Irvine,

653 E. Peltason Dr., Irvine, CA 92697, USA. [3]Klinik für Innere Medizin III, Kardiologie, Angiologie und Internistische Intensivmedizin, Universitätsklinikum des Saarlandes, Homburg/Saar 66421, Germany. [4]Department of Emergency Medicine, Iran University of Medical Sciences, Tehran 14496, Iran. [5]Department of Neurology, Saarland University Medical Center, Homburg/Saar 66421, Germany. [6]Kamyar Clinic, Tehran 51406, Iran. [7]Department of Emergency Medicine, School of Medicine, University of California Irvine, Irvine, CA 92697, USA.

Received: 16 February 2014 Accepted: 30 August 2014
Published online: 26 September 2014

References

1. Boulouffe C, Doucet B, Muschart X, Charlin B, Vanpee D: **Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation.** *Emerg Med J* 2013, **31:**313–316.
2. Humbert AJ, Besinger B, Miech EJ: **Assessing clinical reasoning skills in scenarios of uncertainty: convergent validity for a script concordance test in an emergency medicine clerkship and residency.** *Acad Emerg Med* 2011, **18:**627–634.
3. Carriere B, Gagnon R, Charlin B, Downing S, Bordage G: **Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a script concordance test.** *Ann Emerg Med* 2009, **53:**647–652.
4. Park AJ, Barber MD, Bent AE, Dooley YT, Dancz C, Sutkin G, Jelovsek JE: **Assessment of intraoperative judgment during gynecologic surgery using the script concordance test.** *Am J Obstet Gynecol* 2010, **203**(240)**:**240.e241–246.
5. Mathieu S, Couderc M, Glace B, Tournadre A, Malochet-Guinamand S, Pereira B, Dubost J-J, Soubrier M: **Construction and utilization of a script concordance test as an assessment tool for dcem3 (5th year) medical students in rheumatology.** *BMC Med Educ* 2013, **13:**166.
6. Duggan P, Charlin B: **Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges.** *BMC Med Educ* 2012, **12:**29.
7. Nouh T, Boutros M, Gagnon R, Reid S, Leslie K, Pace D, Pitt D, Walker R, Schiller D, MacLean A, Hameed M, Fata P, Charlin B, Meterissian SH: **The script concordance test as a measure of clinical reasoning: a national validation study.** *Am J Surg* 2012, **203:**530–534.
8. Piovezan RD, Custódio O, Cendoroglo MS, Batista NA, Lubarsky S, Charlin B: **Assessment of undergraduate clinical reasoning in geriatric medicine: application of a script concordance test.** *J Am Geriatr Soc* 2012, **60:**1946–1950.
9. Bursztejn AC, Cuny JF, Adam JL, Sido L, Schmutz JL, de Korwin JD, Latarche C, Braun M, Barbaud A: **Usefulness of the script concordance test in dermatology.** *J Eur Acad Dermatol Venereol* 2011, **25:**1471–1475.
10. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA: **Assessment of clinical reasoning: a script concordance test designed for pre-clinical medical students.** *Med Teach* 2011, **33:**472–477.
11. Kania RE, Verillaud B, Tran H, Gagnon R, Kazitani D, Huy PTB, Herman P, Charlin B: **Online script concordance test for clinical reasoning assessment in otorhinolaryngology: the association between performance and clinical experience.** *Arch Otolaryngol Head Neck Surg* 2011, **137:**751–755.
12. Lambert C, Gagnon R, Nguyen D, Charlin B: **The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning.** *Radiat Oncol* 2009, **4:**7.
13. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B: **The script concordance test: a new tool assessing clinical judgement in neurology.** *Can J Neurol Sci* 2009, **36:**326–331.
14. Meterissian SH: **A novel method of assessing clinical reasoning in surgical residents.** *Surg Innov* 2006, **13:**115–119.
15. Fournier JP, Demeester A, Charlin B: **Script concordance tests: guidelines for construction.** *BMC Med Inform Decis Mak* 2008, **8:**18.
16. Bowen JL: **Educational strategies to promote clinical diagnostic reasoning.** *N Engl J Med* 2006, **355:**2217–2225.
17. Oxman AD, Guyatt GH: **The science of reviewing research.** *Ann N Y Acad Sci* 1993, **703:**125–133. Discussion 133–124.
18. Zehtabchi S, Kline JA: **The art and science of probabilistic decision-making in emergency medicine.** *Acad Emerg Med* 2010, **17:**521–523.
19. Simel DL, Rennie D: *Rational clinical examination: Evidence-based clinical diagnosis.* 1st edition. Chicago: McGraw-Hill; 2009.
20. van Laerhoven H, van der Zaag-Loonen HJ, Derkx BHF: **A comparison of Likert scale and visual analogue scales as response options in children's questionnaires.** *Acta Paediatr* 2004, **93:**830–835.
21. Guyatt GH, Townsend M, Berman LB, Keller JL: **A comparison of Likert and visual analogue scales for measuring change in function.** *J Chronic Dis* 1987, **40:**1129–1133.
22. Straus SERW, Glasziou P, Haynes RB: **Diagnosis and screening**. In *Evidence-based medicine: how to practice and teach EBM.* 3rd edition. London: Elsevier; 2005:67–99.
23. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C: **Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients.** *Br Med J* 1975, **2:**486–489.
24. Petrucci AM, Nouh T, Boutros M, Gagnon R, Meterissian SH: **Assessing clinical judgment using the script concordance test: the importance of using specialty-specific experts to develop the scoring key.** *Am J Surg* 2013, **205:**137–140.
25. Gagnon R, Lubarsky S, Lambert C, Charlin B: **Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither?** *Adv Health Sci Educ Theory Pract* 2011, **16:**601–608.
26. Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, Goudreau J, van der Vleuten C: **Assessment in the context of uncertainty using the script concordance test: more meaning for scores.** *Teach Learn Med* 2010, **22:**180–186.
27. Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauve E, van der Vleuten C: **Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel.** *Med Educ* 2006, **40:**848–854.
28. Straus SE, Glasziou P, Richardson WS, Haynes RB: **Introduction**. In *Evidence-based medicine: How to practice and teach it.* 4th edition. Edinburgh: Churchill Livingstone; 2010:1–12.
29. Ramos K, Linscheid R, Schafer S: **Real-time information-seeking behavior of residency physicians.** *Fam Med* 2003, **35:**257–260.
30. Graber M, Gordon R, Franklin N: **Reducing diagnostic errors in medicine: what's the goal?** *Acad Med* 2002, **77:**981–992.
31. Norman GR, Eva KW: **Diagnostic error and clinical reasoning.** *Med Educ* 2010, **44:**94–100.
32. Nendaz M, Perrier A: **Diagnostic errors and flaws in clinical reasoning: mechanisms and prevention in practice.** *Swiss Med Wkly* 2012, **142:**w13706.
33. Lineberry M, Kreiter CD, Bordage G: **Threats to validity in the use and interpretation of script concordance test scores.** *Med Educ* 2013, **47:**1175–1183.

doi:10.1186/s12245-014-0034-3
Cite this article as: Ahmadi *et al.*: Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios?. *International Journal of Emergency Medicine* 2014 **7**:34.