Infection (2011) 39:135–140
DOI 10.1007/s15010-011-0101-0

CLINICAL AND EPIDEMIOLOGICAL STUDY

Interobserver agreement on signs and symptoms of patients with acute febrile illness

R. P. Daumas · P. Brasil · C. S. Bressan · R. V. C. Oliveira · B. B. G. Carvalho · D. V. Carneiro · S. R. L. Passos

Received: 4 August 2010 / Accepted: 14 February 2011 / Published online: 13 April 2011

Abstract

Purpose To assess the interobserver agreement on clinical history and physical examination when using a semi-structured questionnaire to evaluate patients with an acute febrile illness (AFI).

Methods A cross-sectional study was conducted with outpatients aged 12 years and over, presenting with an AFI defined as fever up to 7 days and no evident focus of infection. Clinical data were collected independently by two physicians using a semi-structured questionnaire. Interobserver agreement was estimated using kappa coefficients with a 95% confidence interval (CI).

Results A total of 140 patients (age range 13–73 years; 56.4% females) were enrolled. All symptoms showed weighted kappa values significantly greater than 0.6, indicating an at least substantial agreement. As most physical signs were infrequent and of mild intensity, they were recoded and analyzed as absent/present. Of the signs with prevalence ≥15%, exanthema, pallor, lymph node enlargement, and eye congestion showed agreements significantly greater than 0.6, while kappa confidence limits for pharyngeal erythema and dehydration included values classified as regular.

Conclusions High agreement was observed for most of the clinical data assessed, and symptom grading was feasible. Some physical findings were rare and their inclusion in a structured form may not be justified in this setting. The questionnaire application showed good reliability for the most frequent signs and symptoms and may prove to be useful at gathering data for surveillance and research at sentinel sites.

Keywords Reproducibility of results · Physical examination · Medical history-taking/methods · Fever · Acute diseases

Introduction

The etiological diagnosis of an acute febrile illness (AFI) depends mainly on the results of microbiological tests, as different diseases have similar clinical findings. Nevertheless, clinical data are essential for the identification of cases thought to have an infectious etiology, since some diseases require specific therapy to be initiated even before diagnostic confirmation by laboratory tests [1].

In developing countries, the high cost of AFI laboratory investigation hinders its wide implementation in health care services. Therefore, the creation of sentinel surveillance sites, capable of identifying infectious agents of high clinical-epidemiological relevance, has been recommended to maintain an active and cost-effective surveillance [2–4]. The analysis of data obtained in such sites could provide the relative frequency of diseases and identify clinical characteristics useful in their differential diagnosis, helping the creation of locally adequate clinical protocols [5].

Electronic supplementary material The online version of this article (doi:10.1007/s15010-011-0101-0) contains supplementary material, which is available to authorized users.

R. P. Daumas (✉) · R. V. C. Oliveira ·
B. B. G. Carvalho · D. V. Carneiro · S. R. L. Passos
Laboratory of Clinical Epidemiology,
Evandro Chagas Clinical Research Institute,
Oswaldo Cruz Foundation, Av Brasil,
4365—Manguinhos, Rio de Janeiro, RJ, Brazil
e-mail: regina.daumas@gmail.com

P. Brasil · C. S. Bressan
Laboratory of Acute Febrile Diseases,
Evandro Chagas Clinical Research Institute,
Oswaldo Cruz Foundation, Av Brasil,
4365—Manguinhos, Rio de Janeiro, RJ, Brazil

Gathering clinical data in a standardized and systematic way enables statistical analyses and the identification of associations between clinical findings, etiology, and prognosis. In this sense, some data collection tools have been developed and evaluated with the aim of assisting the diagnosis of specific conditions based solely on clinical history and examination, such as a structured history-taking on knee problems in the elderly [6] and a reliable and valid questionnaire to detect gastroesophageal reflux disease [7]. However, the evaluation of signs and symptoms of acute febrile patients is not usually done in a standardized way and we could not find structured instruments designed for a comprehensive assessment of these patients.

The diagnostic validity of clinical data can be tested in an analogous way to that recommended for diagnostic laboratory tests [8]. For instance, the accuracy of several clinical signs for detecting pneumonia in children have long been studied [9]. Evaluating the reliability of clinical assessment, defined as the ability of observers to measure a parameter in a reproducible and consistent way, should be the first step of the validation process [10].

In order to standardize the assessment and registration of clinical history and physical examination at a sentinel unit for AFI, a semi-structured questionnaire and a guide for clinical data collection were elaborated and implemented. The aim of this study was to assess the interobserver agreement at identifying and quantifying clinical data when using that tool.

Patients and methods

Development of the questionnaire

A semi-structured questionnaire was elaborated, with questions regarding the presence and intensity of 28 symptoms and 25 clinical signs potentially relevant for the differential diagnosis of AFI.

The definition of the clinical variables of the questionnaire and the classification of their intensity were based on the opinion of infectious disease specialists of a referral center for AFI and on the specialized medical literature [11–13]. For standardized application of the questionnaire, a manual of procedures was elaborated, with precise instructions about the formulation of questions and the criteria to be used in encoding ordinal answers according to their intensity. The data collection tool underwent three pilot tests, followed by improvements in format and content, until its final definition. The clinical team was then trained in using it.

Out of the 28 symptoms assessed, six were collected as dichotomous (present/absent) and 22 as ordinal, the latter being categorized as grades 0, 1, 2, and 3 for absent, mild, moderate, and severe intensity, respectively. The history of fever was categorized according to the highest temperature measured during the disease period as follows: up to 37.5°C; from 37.5 to 38.5°C; and above 38.5°C. The other symptoms, when present, were classified regarding their intensity according to the following criteria: the extent of physical activity limitation; the need for symptom-relief medications and responses to such medications; and the number of episodes (e.g., vomiting, diarrhea) or the extent of body involvement (e.g., rash, petechiae). As a whole, low-intensity symptoms that did not affect physical activity and did not need symptom-relief medications were classified as mild; those moderately affecting physical activity and/or requiring the use of symptom-relief medications were classified as moderate; and those severely affecting physical activity or those that were not relieved with the use of medication were classified as severe. Out of the clinical signs, 11 were collected as dichotomous and 14 as ordinal, and these were classified according to the above-described symptom grading, based on their extension or intensity.

Place and time period of the study

To assess the interobserver agreement for the different items of the protocol, a cross-sectional study was conducted at the following two clinical health care sites in the municipality of Rio de Janeiro, Brazil: the Sentinel Outpatient Clinic for Acute Febrile Illnesses of the Evandro Chagas Clinical Research Institute (IPEC) from the Oswaldo Cruz Foundation (FIOCRUZ) and the Emergency Department of the Lourenço Jorge Municipal Hospital (HMLJ). The inclusion of patients in the study lasted from March 2006 to April 2008 at the Sentinel Outpatient Clinic for Acute Febrile Illnesses of the IPEC-FIOCRUZ, and from October 2007 to January 2008 at the HMLJ.

Criteria of eligibility and selection of patients

Individuals of both genders, more than 12 years of age, presenting with an AFI, defined as a history of fever (anamnestic fever or axillary temperature over 37.0°C) for up to 7 days with no evident focus of infection, were considered to be eligible for the study.

A total of 140 patients were selected by the sequential inclusion of eligible patients during the shifts attended by two physicians of the research team.

The study was approved by the Committee on Ethics in Research of the IPEC-FIOCRUZ, and all participants and/or their guardians provided written informed consent.

Data collection, processing, and analysis

Clinical data were collected during the patient's first medical visit due to the disease, by use of the previously elaborated instrument. Each patient was assessed by two members of the clinical team, which comprised two medical students (MS) and two infectious disease specialists (IDS). Most patients (57%) were evaluated by an IDS–MS pair; 36% by an MS–MS pair; and only 6% by an IDS–IDS pair. During the medical visit, one of the observers asked the questions regarding the presence and intensity of symptoms and both independently recorded the patient's answers. After clinical history-taking, the patient underwent complete physical examination by both observers independently, who recorded the findings in their respective research questionnaires.

Data were entered in an electronic form created with EpiData 3.1 software [14]. Statistical analyses were performed with Stata 9.0 software [15].

The scarcity of findings on physical examination and the low variability in their severity prevented statistical analysis of ordinal agreement. For instance, only four out of 14 clinical signs collected in ordinal format were seen at the maximum intensity: dehydration, exanthema, lymph node enlargement, and petechiae (data not shown). Therefore, mild, moderate, and severe grades were grouped together, and all 25 clinical signs were analyzed as dichotomous variables (absence/presence).

The interobserver agreement was estimated by the use of kappa statistics, which corrects for chance agreement [16]. Simple kappa was calculated for dichotomous variables and quadratic weighted kappa for ordinal variables. The 95% confidence intervals for kappa were estimated by the use of the analytical method in the case of dichotomous variables [17], and by the use of bootstrapping in the case of ordinal variables [18]. Kappa values were interpreted according to the Landis and Koch classification as follows: no agreement ($k < 0.0$); slight agreement ($0.0 < k < 0.2$); fair agreement ($0.2 < k < 0.4$); moderate agreement ($0.4 < k < 0.6$); substantial agreement ($0.6 < k < 0.8$); almost perfect agreement ($0.8 < k < 1.00$); and perfect agreement ($k = 1$) [19]. For signs and symptoms with prevalences ranging from 15 to 85%, this study sample was sufficient to estimate kappa values greater than 0.80, with 0.15 absolute error, and 95% confidence limits.

To provide a better interpretation of the results, the proportions of positive (P_{pos}) and negative (P_{neg}) agreement were also presented for dichotomous variables [20] (Fig. 1).

Results

This study assessed 140 patients, 56.4% of whom were females. The age of the patients ranged from 13 to 73 years (mean age = 34 years; standard deviation [SD] = 13.5). The time elapsed from reported fever onset until medical visit ranged from 0 to 7 days (median = 3 days).

		Observer 1		
		Yes	No	*Total*
Observer 2	Yes	a	b	*g1*
	No	c	d	*g2*
	Total	*f1*	*f2*	*N*

$$Po\ (observed\ agreement) = \frac{a+d}{N}$$

$$Pe\ (chance\ agreement) = \frac{f1 \times g1 + f2 \times g2}{N}$$

$$\text{kappa} = \frac{Po - Pe}{1 - Pe}$$

$$\text{P}_{\text{pos}} = \frac{2a}{f1 + g1} \quad ; \qquad \text{P}_{\text{neg}} = \frac{2d}{f2 + g2}$$

Fig. 1 Formulae used to calculate agreement rates for dichotomous variables according to the classification of patients by observers 1 and 2

Table 1 presents the frequency and weighted kappa value for each of the 22 symptoms assessed regarding their presence and intensity and classified into four categories (absent, mild, moderate, and severe). These symptoms showed weighted kappa values greater than 0.75, indicating agreements from substantial to almost perfect. Although all patients reported having had fever, only 84 (60.0%) had measured their body temperature. The report of the highest temperature measured by the patient during the disease period showed substantial reliability, but the precision of that estimate was limited (kappa = 0.780; 95% confidence interval [CI]: 0.480–0.947).

Table 2 shows the frequency, P_{pos} and P_{neg}, total observed agreement, and the simple kappa value for the symptoms assessed regarding their presence, with no classification of intensity. All symptoms showed kappa values greater than 0.6, indicating an at least substantial agreement. Of those symptoms, choluria had the lowest kappa value (0.685; 95% CI: 0.484–0.886), as well as the lowest prevalence and mean positive agreement values.

All clinical signs were classified into two categories (absent/present). Table 3 shows the agreement rates for 12 clinical signs with frequency greater than 5%. Most signs identified on physical examination had low frequency and mild intensity. Only six signs had frequency greater than 15%, making possible the calculation of the simple kappa with the aimed precision. For those signs, the kappa values indicated agreements ranging from substantial to almost perfect, and the lowest limits of the 95% confidence interval were greater than 0.6 for all signs, except for pharyngeal erythema (0.619; 95% CI: 0.464–0.773) and

Table 1 Interobserver agreement for symptoms assessed and classified as absent, mild, moderate, or intense (0–3)

	n	Mean prevalence (%)	Weighted kappa (95% confidence interval [CI])
Measured fever	84	100.0	0.780 (0.480–0.947)
Exhaustion	138	95.6	0.837 (0.750–0.901)
Myalgia	139	90.0	0.805 (0.679–0.900)
Headache	140	89.3	0.888 (0.827–0.930)
Anorexia	138	83.3	0.868 (0.774–0.925)
Lumbar pain	138	83.3	0.915 (0.851–0.957)
Chills	140	64.6	0.869 (0.786–0.931)
Retro-orbital pain	138	64.1	0.878 (0.793–0.931)
Nausea	139	62.6	0.936 (0.898–0.965)
Arthralgia	140	60.7	0.932 (0.887–0.964)
Photophobia	138	51.8	0.917 (0.867–0.958)
Dizziness	139	45.0	0.811 (0.726–0.875)
Rash	140	37.8	0.843 (0.739–0.916)
Abdominal pain	139	36.0	0.970 (0.943–0.990)
Vomiting	138	32.6	0.966 (0.938–0.988)
Itching	140	31.8	0.755 (0.607–0.876)
Sore throat	140	28.2	0.886 (0.784–0.960)
Dry cough	137	26.3	0.951 (0.908–0.984)
Dyspnea	139	15.8	0.878 (0.749–0.951)
Hoarseness	140	15.0	0.846 (0.656–0.966)
Earache	139	11.5	0.825 (0.594–0.951)
Productive cough	138	10.9	0.896 (0.729–0.990)

Table 2 Interobserver agreement for symptoms assessed as present/absent in patients with acute febrile illness (AFI)

Symptoms	*n*	Mean prevalence (%)	P_{pos}	P_{neg}	Po	Simple kappa (95% CI)
Taste disorder	139	68.0	0.974	0.944	0.964	0.917 (0.846–0.988)
Coryza	71	25.9	0.889	0.961	0.942	0.850 (0.750–0.951)
Diarrhea	140	31.1	0.989	0.995	0.993	0.983 (0.951–1.000)
History of bleeding[a]	139	20.5	0.912	0.977	0.964	0.890 (0.795–0.984)
Nasal congestion	139	20.1	0.821	0.955	0.928	0.777 (0.646–0.908)
Choluria	140	10.0	0.714	0.968	0.943	0.685 (0.484–0.886)

P_{pos}, mean positive agreement proportion; P_{neg}, mean negative agreement proportion; Po, total observed agreement proportion

[a] Bleeding sites: metrorrhagia (40%), gingival bleeding (23.3%), epistaxis (20%), melena (10%), hemoptysis (10%), hematuria (3.3%), and hematemesis (3.3%)

dehydration (0.718; 95% CI: 0.562–0.874). Signs of enanthema and edema had kappa indices classified as fair and moderate, respectively. When analyzing the other agreement rates, a low P_{pos} was observed for those signs (equal to or lower than 0.50 for both).

No kappa value was calculated for signs whose frequency was lower than 5%, such as those related to shock (cold extremities, delayed capillary refill, hypotension, and filiform pulse), alterations in the respiratory system (sibilus, rales, crackles, and labored breathing), hemorrhage signs other than petechiae (purpura, gingival bleeding), lumbar percussion pain, neck stiffness, and ascitis.

Discussion

In this study, high reliability indices were observed for most of the clinical data assessed. Regarding accuracy, the ordinal collection of symptoms by the use of the standardized questionnaire was successful. The fact that most

Table 3 Prevalence, positive and negative agreement, total observed agreement, and simple kappa for clinical signs in patients with AFI

Clinical signs	*n*	Mean prevalence (%)	P_{pos}	P_{neg}	Po	Simple kappa (95% CI)
Exanthema[a]	136	55.2	0.973	0.967	0.971	0.941 (0.883–0.998)
Pallor	137	36.1	0.828	0.903	0.876	0.732 (0.615–0.850)
Lymph node enlargement	136	36.0	0.878	0.931	0.912	0.809 (0.706–0 .912)
Eye congestion	136	26.9	0.806	0.929	0.896	0.735 (0.606–0.864)
Pharyngeal erythema	130	25.8	0.716	0.902	0.854	0.619 (0.464–0.773)
Dehydration	137	17.2	0.766	0.952	0.920	0.718 (0.562–0.874)
Hepatomegaly	132	8.7	0.870	0.988	0.977	0.857 (0.699–1.000)
Petechiae	137	8.4	0.870	0.988	0.978	0.858 (0.700–1.000)
Enanthema	137	6.9	0.421	0.957	0.920	0.388 (0.109–0.668)
Edemas	138	5.8	0.500	0.969	0.942	0.469 (0.157–0.782)
Cardiac murmur	136	5.6	0.800	0.988	0.977	0.789 (0.557–1.000)
Splenomegaly	135	5.2	0.857	0.992	0.985	0.849 (0.644–1.000)

P_{pos}, mean positive agreement proportion; P_{neg}, mean negative agreement proportion; Po, total observed agreement proportion

[a] Classified as macular (68/78) or macular–papular (10/78)

pairs included at least one medical student and not only specialists suggests that the results obtained can be achieved with some training in the use of the protocol and are not limited to experts. Considering that the differential diagnosis of AFI and the notification of suspicious cases are based on combinations of those signs and symptoms, structured clinical history taken in sentinel units enables the acquisition of information useful for epidemiological surveillance. For instance, elevated agreement rates were found for the presence of exanthema ($k = 0.941$; 0.883–0.998) and history of itching ($k = 0.755$; 0.607–0.876), two clinical signs previously identified as useful for dengue diagnosis in the state of Rio de Janeiro [21]. It is worth noting that the possibility of collecting more detailed data, comprising information on intensity grading, is potentially innovative for clinical research in the field.

It should be emphasized that, for most signs and symptoms, there is no gold-standard test and the assessment of reliability may be the main parameter to evaluate validity. However, because of the acute and dynamic character of AFI, test–retest reliability could not be assessed. A clinical steady state cannot be assumed for more than 24 h and repeating clinical interview on the same day would have considerable recall effect.

We found relatively high interobserver agreement when compared with other published studies. For instance, we achieved much higher kappa values for pharyngeal erythema (0.62 vs. 0.17) and history of cough (0.95 vs. 0.70) than Schwartz et al. [22]. This may be due to the fact that we worked with a small and well-trained clinical team, while they evaluated primary care physicians in routine clinical practice.

A significant number of patients did not measure their temperature with a thermometer. This represents an additional difficulty in the initial assessment and follow-up of cases suspected of an AFI, because the level of fever and its pattern of occurrence are important for the differential diagnosis of febrile diseases. However, as it is very common among users of public health services in Brazil not to measure body temperature when feeling feverish, we found it better to include patients with anamnestic fever than to exclude all those who had not used a thermometer. Although some patients with anamnestic fever could not have true fever, the interobserver agreement on signs and symptoms was probably not affected by this issue.

Abnormal physical findings were infrequent and of mild intensity, hindering the assessment of interobserver agreement regarding intensity. Considering that the same limitation will impair other quantitative analyses, the effort to grade the intensity of these findings may be unjustified in this setting. The usual recording of the presence or absence of clinical signs could be adopted with no significant loss for analysis. Signs that are already collected as ordinal in the clinical practice should remain as such (e.g., dehydration, pallor, jaundice).

Some clinical signs related to more severe disease had extremely low frequency, preventing any reliability statistical analysis. Those signs should be studied in a hospital setting.

The multiplicity of clinical situations faced by a physician in the context of an AFI clinic cannot be fully encompassed in any single set of closed questions, no matter how comprehensive they are. The attempt to include a very large set of information would result in an unfeasible tool and disrupt clinical reasoning. For example, when assessing a single symptom, the formulation of diagnostic hypotheses and/or their rejection can require information regarding the symptom's onset, intensity, duration, frequency, quality, context, location, relief, and worsening factors [23]. This kind of essential information cannot be anticipated in a standardized form.

The instrument evaluated in this study is not intended to replace regular anamnesis, individualized clinical assessment, and registration on medical records. Instead, it should be viewed as a complementary tool, aimed at recording clinical data in such a way that they are useful for surveillance and research. As such, it should contain questions about common findings in the target population that are potentially discriminating for the diagnosis and/or useful for surveillance. This parsimonious approach towards data collection and their characterization avoids disturbing routine health care and can promote better acceptance.

Based on our results, we could propose some changes to the initial instrument in order to simplify data collection and optimize reliability. The analysis of less reliable items was useful to identify sources of interobserver variability, contributing to enhance the initial tool and the manual of procedures. Extremely rare signs should be omitted from the instrument and, instead, be registered on medical records in an ordinary fashion.

The questionnaire application showed good reliability for the most frequent signs and symptoms of acute febrile patients and may prove to be useful at gathering data for surveillance and research at sentinel sites. The accuracy of these signs and symptoms to diagnose specific febrile diseases, such as dengue and malaria, could be investigated in endemic areas by using this tool.

Acknowledgments We are grateful to: Liliane Alves Reis and Marlene da Conceição Rezende for the administrative support; Patrícia Rosana de Souza for contribution to the data collection; João Claudio Arnaldo Alves e Flávia Lattario Ribeiro for help with the data entry; emergency medical staff, laboratory personnel, epidemiology service team, and the director of the Lourenço Jorge Municipal Hospital for all of the support provided. This work was supported by grants from the Fundação Oswaldo Cruz, Programa de Desenvolvimento e Inovação Tecnológica em Saúde Pública (PDTSP-Dengue—RDCL 05 Project) and Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq (MS-SCTIE-DECIT 25/2006).

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Dassanayake DL, Wimalaratna H, Agampodi SB, Liyanapathirana VC, Piyarathna TA, Goonapienuwala BL. Evaluation of surveillance case definition in the diagnosis of leptospirosis, using the microscopic agglutination test: a validation study. BMC Infect Dis. 2009;9:48.
2. Crump JA, Youssef FG, Luby SP, Wasfy MO, Rangel JM, Taalat M, et al. Estimating the incidence of typhoid fever and other febrile illnesses in developing countries. Emerg Infect Dis. 2003;9:539–44.
3. Gubler DJ. Surveillance for dengue and dengue hemorrhagic fever. Bull Pan Am Health Organ. 1989;23:397–404.
4. Marzochi KBF. Endemic dengue: surveillance strategy challenges. Rev Soc Bras Med Trop. 2004;37:413–5. doi:10.1590/S0037-86822004000500009.
5. Archibald LK, Reller LB. Clinical microbiology in developing countries. Emerg Infect Dis. 2001;7:302–5.
6. Peat G, Wood L, Wilkie R, Thomas E. How reliable is structured clinical history-taking in older adults with knee problems? Inter- and intraobserver variability of the KNE-SCI. J Clin Epidemiol. 2003;56:1030–7. doi:10.1016/S0895-4356(03)00204-X.
7. Manterola C, Muñoz S, Grande L, Bustos L. Initial validation of a questionnaire for detecting gastroesophageal reflux disease in epidemiological settings. J Clin Epidemiol. 2002;55:1041–5. doi:10.1016/S0895-4356(02)00454-7.
8. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Fam Pract. 2004;21:4–10.
9. Margolis P, Gadomski A. The rational clinical examination. Does this infant have pneumonia? JAMA. 1998;279:308–13. doi:10.1001/jama.279.4.308.
10. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.
11. Porto CC. Semiologia médica. 5th ed. Rio de Janeiro: Guanabara Koogan; 2005.
12. Schechter M, Marangoni DV. Doenças infecciosas: conduta diagnóstica e terapêutica. 2nd ed. Rio de Janeiro: Guanabara Koogan; 1998.
13. Kasper DL, Braunwald E, Hauser S, Longo DL, Jameson JL, Fauci AS. Harrison's principles of internal medicine. 16th ed. New York: McGraw-Hill Medical Publishing Division; 2005.
14. Lauritsen JM, Bruus M. EpiData (version 3). A comprehensive tool for validated entry and documentation of data. Odense, Denmark: The EpiData Association; 2003–2004.
15. StataCorp. Stata Statistical Software: Release 9. College Station, TX: StataCorp LP; 2005.
16. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
17. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981.
18. Reichenheim ME. Confidence intervals for the kappa statistic. Stata J. 2004;4:421–8.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
20. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43:551–8. doi:10.1016/0895-4356(90)90159-M.
21. Passos SR, Bedoya SJ, Hökerberg YH, Maia SC, Georg I, Nogueira RM, et al. Clinical and laboratory signs as dengue markers during an outbreak in Rio de Janeiro. Infection. 2008;36:570–4. doi:10.1007/s15010-008-7334-6.
22. Schwartz K, Monsur J, Northrup J, West P, Neale AV. Pharyngitis clinical prediction rules: effect of interobserver agreement: a MetroNet study. J Clin Epidemiol. 2004;57:142–6. doi:10.1016/S0895-4356(03)00249-X.
23. Takemura Y, Atsumi R, Tsuda T. Identifying medical interview behaviors that best elicit information from patients in clinical practice. Tohoku J Exp Med. 2007;213:121–7. doi:10.1620/tjem.213.121.