Journal of Clinical Epidemiology 112 (2019) 20−27

ORIGINAL ARTICLE

Expert panel diagnosis demonstrated high reproducibility as reference standard in infectious diseases

Chantal B. van Houten[a], Christiana A. Naaktgeboren[b], Liat Ashkenazi-Hoffnung[c], Shai Ashkenazi[d], Wim Avis[e], Irena Chistyakov[f], Teresa Corigliano[g], Annick Galetto[g], Iker Gangoiti[h], Alain Gervaix[g], Daniel Glikman[i], Inga Ivaskeviciene[j], Amir A. Kuperman[k], Laurence Lacroix[g], Yvette Loeffen[a], Fanny Luterbacher[g], Clemens B. Meijssen[l], Santiago Mintegi[h], Basheer Nasrallah[f], Cihan Papan[m], Annemarie M.C. van Rossum[n], Henriette Rudolph[l], Michal Stein[o], Roie Tal[p], Tobias Tenenbaum[l], Vytautas Usonis[j], Wouter de Waal[q], Stefan Weichert[l], Joanne G. Wildenbeest[a], Karin M. de Winter-de Groot[r], Tom F.W. Wolfs[a], Niv Mastboim[s], Tanya M. Gottlieb[s], Asi Cohen[s], Kfir Oved[s], Eran Eden[s], Paul D. Feigin[t], Liran Shani[s], Louis J. Bont[a,*], the IMPRIND consortium

[a]*Division of Pediatric Immunology and Infectious Diseases, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands*
[b]*Division Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, The Netherlands*
[c]*Schneider Children's Medical Center, Petach Tikva, Sackler Faculty of Medicine, Tel Aviv, Israel*
[d]*Adelson School of Medicine, Ariel University, Schneider Children's Medical Center, Petach Tikva, Israel*
[e]*Department of Pediatrics, Gelderse Vallei Hospital, Ede, The Netherlands*
[f]*Department of Pediatrics, Bnai Zion Medical Centre, Haifa, Israel*
[g]*Department of Pediatrics, Geneva University Hospitals, Geneva, Switzerland*
[h]*Department of Pediatric Emergency Medicine, Cruces University Hospital, Bilbao, Spain*
[i]*Infectious Diseases Unit, Padeh Poria Medical Center and the Azrieli faculty of Medicine in the Galilee, Bar-Ilan University, Safed, Israel*
[j]*Clinic of Children Diseases, Institute of Clinical medicine, Faculty of Medicine, Vilnius University Vilnius, Lithuania*
[k]*Blood Coagulation Service and Pediatric Hematology Clinic, Galilee Medical Centre, Nahariya, and Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel*
[l]*Department of Pediatrics, Meander Medical Centre, Amersfoort, The Netherlands*
[m]*Pediatric Infectious Diseases, University Children's Hospital Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[n]*Department of Pediatrics, Erasmus MC University, Medical Centre Rotterdam, The Netherlands*
[o]*Department of Pediatrics, Hillel Yaffe Medical Centre, Hadera, Israel*
[p]*Department of Pediatrics, Galilee Medical Centre, Nahariya and Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel*
[q]*Department of Pediatrics, Diakonessenhuis, Utrecht, The Netherlands*
[r]*Department of Pediatric Respiratory Medicine, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands*
[s]*MeMed, Tirat Carmel, Israel*
[t]*Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa, Israel*

Accepted 18 March 2019; Published online 28 March 2019

Abstract

Objective: If a gold standard is lacking in a diagnostic test accuracy study, expert diagnosis is frequently used as reference standard. However, interobserver and intraobserver agreements are imperfect. The aim of this study was to quantify the reproducibility of a panel diagnosis for pediatric infectious diseases.

Study Design and Setting: Pediatricians from six countries adjudicated a diagnosis (i.e., bacterial infection, viral infection, or indeterminate) for febrile children. Diagnosis was reached when the majority of panel members came to the same diagnosis, leaving others

Funding: The first author (C.H.) was funded by the TAILORED-Treatment study, which received funding from the EU's Seventh Framework Program FP7 under REA grant agreement No. HEALTH--F3−602860-2013 (TAILORED-Treatment; www.tailored-treatment.eu/).

Conflict of interest statement: The authors have no conflicts of interest relevant to this article to disclose.

Authors' contributions: All authors were responsible for the design of the study. C.H., L.A., S.A., W.A., I.C., T.C., A.G., I.G., A.G., D.G., I.I., A.K., L.L., Y.L., F.L., C.M., S.M., B.N., C.P., A.R., H.R., M.S., R.T., T.T., V.U., W.W., S.W., J.W., K.W., T.W., N.M., L.S., and L.B. performed the experiments. All authors interpreted the data. C.H., C.N., N.M., T.G., A.C., K.O., E.E., P.F., L.S., and L.B. analyzed the data. All authors were responsible for writing and/or critically revisions of the article.

* Corresponding author: University Medical Centre Utrecht, Pediatric Immunology and Infectious Diseases, P.O. Box 85090, Office KC.03.063.0, 3508 AB, Utrecht, The Netherlands. Tel.: +31 88 75 540 03; fax: +31 88 75 553 50.

E-mail address: l.bont@umcutrecht.nl (L.J. Bont).

https://doi.org/10.1016/j.jclinepi.2019.03.010

inconclusive. We evaluated intraobserver and intrapanel agreement with 6 weeks and 3 years' time intervals. We calculated the proportion of inconclusive diagnosis for a three-, five-, and seven-expert panel.

Results: For both time intervals (i.e., 6 weeks and 3 years), intrapanel agreement was higher (kappa 0.88, 95%CI: 0.81-0.94 and 0.80, 95%CI: NA) compared to intraobserver agreement (kappa 0.77, 95%CI: 0.71-0.83 and 0.65, 95%CI: 0.52-0.78). After expanding the three-expert panel to five or seven experts, the proportion of inconclusive diagnoses (11%) remained the same.

Conclusion: A panel consisting of three experts provides more reproducible diagnoses than an individual expert in children with lower respiratory tract infection or fever without source. Increasing the size of a panel beyond three experts has no major advantage for diagnosis reproducibility.

Keywords: Reference standard; Gold standard; Diagnosis; Expert panel; Infectious diseases; Reproducibility

1. Background

Introduction of novel diagnostics into routine care is usually based on clinical validation studies. Evidence on the diagnostic accuracy is produced by comparing results of the test under evaluation (the index test, e.g., a newly developed blood test) with the actual presence or absence of a target condition. Ideally, a gold standard to determine the presence of this target condition is available, which is an error-free classification in all patients, blinded from the index test result, and performed within a short interval of time [1]. There are, however, many conditions, such as certain infectious diseases [2], for which such a gold standard does not exist. Using expert diagnosis as reference standard is therefore commonly applied in validation studies. The experts identify those with the target condition among the persons being tested, based on the available information [1,3]. As a gold standard to diagnose bacterial infections is lacking, it is not possible to analyze the validity (e.g., the accuracy) of these experts. Analyzing the reproducibility of expert diagnoses is therefore most informative. Several studies have assessed agreement between different experts and found a poor to moderate interobserver agreement [4–10]. Because of imperfect reproducibility of the expert diagnosis, the estimated accuracy of the diagnostic test under evaluation may be an overestimation or underestimation [11,12]. Based on these results, regulators and clinicians will decide to use novel diagnostics in clinical practice or not. To overcome these limitations, expert panels have been employed. However, to our knowledge, the reproducibility of such panel diagnoses in infectious diseases has not been examined. The present study was designed to quantify the reproducibility of an expert panel diagnosis for infectious diseases during childhood. This study had two primary objectives to assess reproducibility of an expert panel diagnosis. The first objective was to evaluate the reproducibility of an expert panel diagnosis in comparison to a single expert diagnosis. For this objective, we calculated intraobserver and intrapanel agreement, comparing the diagnoses changed after 6 weeks and 3 years' time intervals. The second objective was to evaluate the proportion of children in whom diagnosis changed after expanding the panel size (i.e., three, five, or seven experts). We included uneven numbers of panel sizes as this is most commonly used in diagnostic studies [13].

2. Methods

For this study, we used data from our previously published OPPORTUNITY study [14]. Briefly, we validated a host-protein–based diagnostic using data from children aged 1 to 60 months presenting with lower respiratory tract infections or fever without source at the emergency department or on pediatric wards in the Netherlands and Israel [14]. Healthy controls were excluded for the present study.

2.1. Participating experts

The Improving diagnostics in infectious diseases (IMPRIND) consortium consisted of 25 pediatricians from six different European countries with at least 10 years of clinical experience since accredited as medical doctors. All experts were blinded to the diagnoses of their peers. Panel members did not receive other training than an explanation of the adjudication process and a review of pediatric, emergency medicine, and infectious disease literature written by the study team as background information without any instructions or decision rules. Before reviewing the cases, experts completed a short online survey on baseline characteristics (e.g., age, hours of clinical work per week).

2.2. Reviewing process

The review process and the definitions employed have been previously described [14]. In short, each expert received an electronic case record form (eCRF) including demographics, medical history, physical examination findings, all available radiological, and laboratory data collected for routine care, microbiological investigation results (including study-specific nasal swab data), and a 28-day follow-up information. Based on the eCRF data, each expert assigned one of the following etiologies to each case: bacterial infection, viral infection, mixed infection (i.e., bacterial and viral coinfection), noninfectious disease, or indeterminate. Cases assigned as mixed infection were later classified as bacterial because both require antimicrobial treatment. Cases were assigned an inconclusive panel diagnosis if each panel member assigned a different diagnosis or when at least the majority of the panel members diagnosed the case as indeterminate.

What is new?

Key findings

- A panel consisting of three experts provided more reproducible diagnoses than an individual expert.
- Increasing the size of a three-expert panel to five or seven experts has no major advantage.

What this adds to what was known?

- This is the first study providing insights into reproducibility of an expert panel diagnosis compared to a single expert diagnosis for infectious diseases during childhood.

What is the implication and what should change now?

- A panel diagnosis of experts, instead of the diagnosis of a single expert, is highly preferred as a reference standard.

2.3. Sample size consideration

Sample size consideration was based on examples from the literature and logistically feasibility (e.g., the number of experts that was still available after a 3-year time interval). Two articles that have compared single observer diagnoses with panel diagnoses on dementia and seizures included 45 and 100 cases, respectively [15,16]. Our first objective was also to evaluate the reproducibility of an expert panel diagnosis in comparison to a single expert diagnosis. We therefore aimed to have at least 100 cases re-evaluated by a panel. For the second objective, we aimed to have at least 50 cases to be diagnosed by panels of seven experts. Each expert reviewed two batches of up to 50 cases (i.e., in total there were a maximum of 100 patient records per expert).

2.4. Case selection

In total, 611 cases from the OPPORTUNITY study were available for the present study. Overall, case selection was based on the sample size consideration as described earlier in this article. For the first objective, evaluating the reproducibility of an expert panel diagnosis in comparison to a single expert diagnosis, we used two different time intervals (i.e., 6 weeks and 3 years, Fig. 1). Randomly selected cases were used to calculate intraobserver agreement. Random selection of the cases was performed using an online sequence generator (https://www.random.org/sequences/). After 6 weeks or 3 years, the same cases were diagnosed by the same experts. The experts received also the same eCRF information (including information from the 28-day follow-up assessment). To quantify reproducibility of panel diagnosis, a random selection of cases was diagnosed by three experts

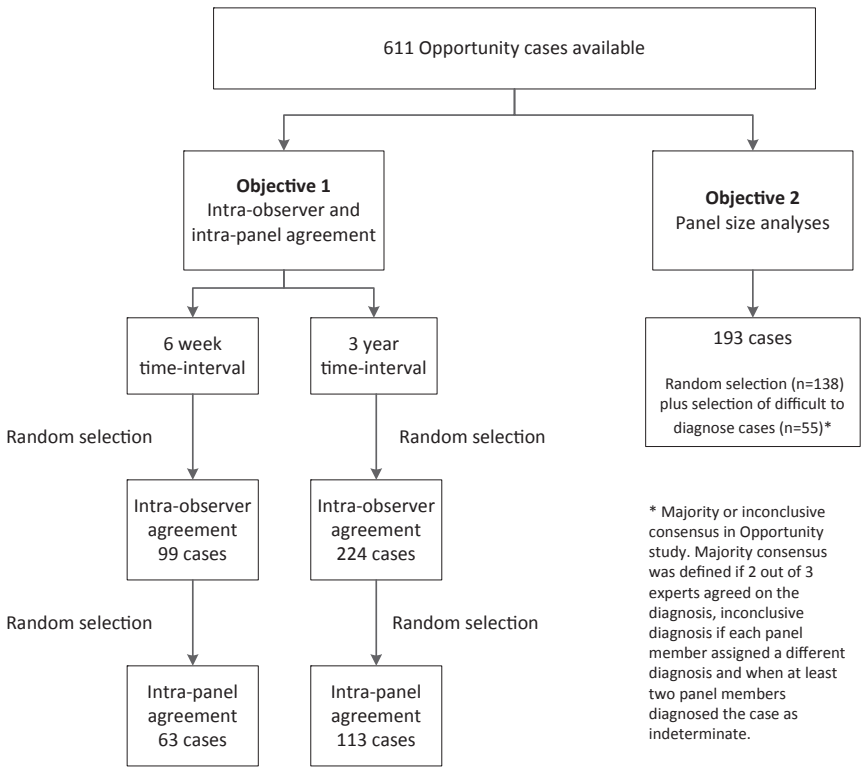

Fig. 1. Case selection. Case selection to assess reproducibility of panel diagnoses. Objective 1: to compare reproducibility of individual expert vs. expert panel diagnosis. Objective 2: to compare diagnosis and agreement for expert panels of different size (i.e., three, five, or seven experts). Case selection was based on a convenient sample size.

and was used to calculate intrapanel agreement (Figure 1). To minimize the possibility that experts had recognized cases from their first assignment (i.e., 6 weeks or 3 years ago), we did not inform the experts about the objective to measure reproducibility. For the second objective, i.e., effect of panel size on reproducibility, case selection was performed in two ways (Figure 1). First, we took a random selection of available cases. In the OPPORTUNITY study, 70% of the cases were assigned an unanimous panel diagnosis. The need for larger expert panels is highest in harder to diagnose cases. We aimed to make our data most suitable for these situations. Therefore, we included a purposive sample of harder to diagnose cases. Harder to diagnose cases included cases with a majority (2 of 3 experts agreed on the diagnosis) or inconclusive diagnosis, thus excluding cases with unanimous diagnosis [14]. The selected cases were reviewed by at least seven experts allowing analyses of different panel compositions.

2.5. Statistical analysis

Intraobserver and interobserver and intrapanel and interpanel agreements were calculated with Fleiss kappa statistics (κ). Kappa statistics measure the observed level of agreement between diagnoses for a set of nominal ratings and corrects for agreement that would be expected by chance [17]. The degree of agreement was rated as suggested by Landis and Koch: fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), or almost perfect (0.81-1.0) [18]. The value of kappa statistics is a relative measure, whereas clinician's question of reference standard variation calls for an absolute measure [11]. Therefore, we also calculated absolute agreement. In addition, we calculated the proportion of bacterial, viral, and inconclusive diagnoses and the alterations in diagnosis for all possible panel compositions if the panel size increased (i.e., three, five, or seven experts). To analyze the reproducibility of different panel sizes, we calculated interobserver and interpanel agreement for all possible panel compositions (e.g., 193 patients × 35 possible panels when constructing a panel with three members from seven options). Analyses of interpanel agreement was performed for three- and five-expert panels; each panel consists of three and five unique panel members. When calculating the proportion of bacterial, viral, and inconclusive diagnoses, and the intrapanel and interpanel agreement, the majority diagnosis was used. Majority diagnosis was defined as at least two out of three, three out of five, or five out of seven panel members agreeing on a diagnosis. We used R version 3.4.4 for the statistical analysis.

3. Results

3.1. Experts and cases characteristics

Twenty-five pediatricians from six different countries participated in the study, of who the majority had special interest in infectious diseases (Appendix Table A1). Thirteen experts (52%) had participated in an expert panel for a clinical study in the past. The experts made a total of

Table 1. Baseline characteristics of the included cases.

Characteristics	Objective 1 (n = 296)	Objective 2 (n = 193)
Age, months, median [IQR]	14 [7-30]	18 [10-31]
Gender, male, *n* (%)	175 (59%)	117 (61%)
Max. temperature, °C, median [IQR]	39.3 [38.8-40.0]	39.5 [39.0-40.0]
Duration of symptoms, d, median [IQR]	2 [1-3]	2 [1-4]
Hospital admission, *n* (%)	185 (63%)	127 (66%)
Hospitalization duration, d, median [IQR]	3 [2-5]	3 [2-5]
Antibiotics prescribed, *n* (%)	164 (55%)	108 [56%]
CRP (mg/L), median [IQR]	21 [7-58]	27 [10-62]
Need of mechanical ventilation, *n* (%)	2 (1%)	3 (2%)
Recruiting site		
Secondary care center, *n* (%)	260 (88%)	169 (87)
Tertiary care center, *n* (%)	33 (11%)	19 (10%)
PICU, *n* (%)	3 (1%)	5 (3%)
Clinical syndrome		
URTI, *n* (%)	89 (30%)	57 (29%)
FWS, *n* (%)	75 (25%)	46 (24%)
LRTI, *n* (%)	66 (22%)	55 (29%)
Other, *n* (%)	66 (22%)	35 (18%)

Abbreviations: IQR, interquartile range; CRP, C-reactive protein; PICU, pediatric intensive care unit; FWS, fever without source; URTI, upper respiratory tract infection; LRTI, lower respiratory tract infection.

Objective 1: to compare reproducibility of individual expert vs. expert panel diagnosis. Objective 2: to compare diagnosis and agreement for expert panels of different size (i.e., three, five, or seven experts).

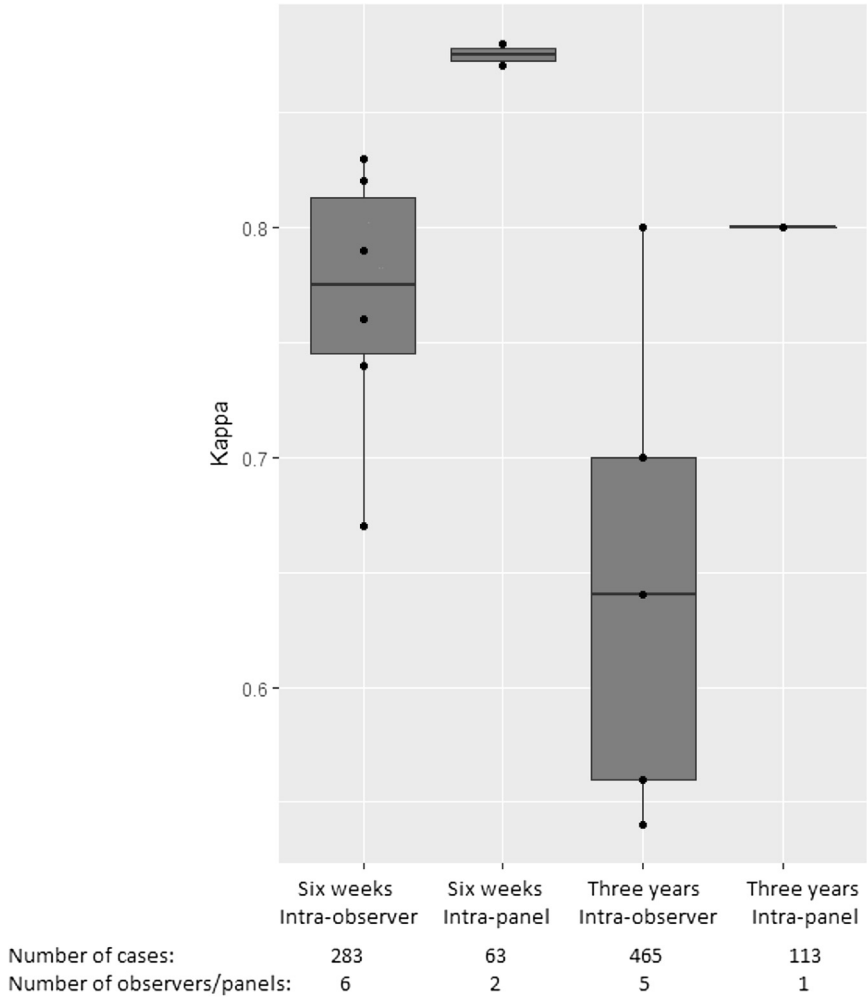

Fig. 2. Box plot for the intraobserver and intrapanel agreements. Each dot corresponds to the kappa per expert or per three-expert panel. In box plots, the band inside the box represents the median kappa. The bottom and top of the box represent the first and third quartiles. The whiskers reflect the minimum and maximum kappa.

2375 individual diagnoses. In total, 355 unique cases were used with some cases being used more than once for the different objectives. The median age of cases used for objective 1 (i.e., intraobserver vs. intrapanel agreement) was 14 months (Table 1). Cases used for objective 2 (i.e., reproducibility for different panel sizes) were slightly older (median age 18 months). In both groups, approximately 60% were male (Table 1).

3.2. Intraobserver and intrapanel agreement

As shown in Figure 2, when the time interval was 6 weeks, intraobserver agreement was substantial (mean $\kappa = 0.77$, 95% CI: 0.71-0.83), when using Landis' cutoff values. When a three-expert panel was employed, with the diagnosis defined by majority, reproducibility was higher with an almost perfect agreement (mean $\kappa = 0.88$, 95% CI: 0.81-0.94). Intrapanel agreement with a 3-year time interval was also higher in comparison to intraobserver agreement (mean $\kappa = 0.80$, 95% CI: NA and mean $\kappa = 0.65$, 95% CI: 0.52-0.78, respectively). Similar to kappa statistics, the absolute proportion of intrapanel agreement was higher compared to the intraobserver agreement for both 6 weeks and 3 years' time intervals (Table 2). Intraobserver agreement as well as intrapanel agreement was modestly higher when the time interval between assessments was shorter (6 weeks vs. 3 years, Table 2). We constructed reclassification tables to obtain deeper insight into diagnostic changes (Appendix Table A2). With a 6-week's time interval, single expert diagnoses changed in 10% of the individual cases and panel diagnoses in 5% of the cases. For a 3-year's time interval, these percentages were 15% and 10%, respectively.

3.3. Panel size

The overall proportion of inconclusive diagnoses for a panel of three experts, using majority consensus, was 11%; this proportion did not decrease when experts were added to the panel (Fig. 3). Moreover, the overall proportion of bacterial and viral diagnoses did not change when the panel was

Table 2. Absolute proportion of intraobserver and intrapanel agreements.

Agreement	% Agreement [min-max]	Number of expert (or panels)
Six weeks interval		
Intraobserver agreement, mean [min-max]	91% [85-94%]	6 experts
Intrapanel agreement, mean [min-max]	95% [94-97%]	2 panels
Three years interval		
Intraobserver agreement, mean [min-max]	84% [76-92%]	6 experts
Intrapanel agreement	90%	1 panel

Presented percentages are mean proportions with minimum and maximum values from different experts or three-expert panels.

expanded to five members. Nevertheless, the diagnosis changed in 11% of the individual cases when the size of the panel increased from three to five experts and in 10% when expanding the panel size from five to seven experts. Interobserver agreement within a panel of three experts was fair with a mean kappa of 0.39 (95% CI: 0.38-0.40). Interobserver agreement did not improve when the panel was expanded to five experts (Appendix Figure A1). Interpanel agreement between different panels was substantial ($\kappa = 0.75$, 95% CI: 0.74-0.75) using the majority diagnosis. Interpanel agreement was higher ($\kappa = 0.86$, 95% CI: 0.85-0.86) when the three-expert panels were expanded to five experts (majority diagnosis defined as at least three out of five panel members agreed on a diagnosis).

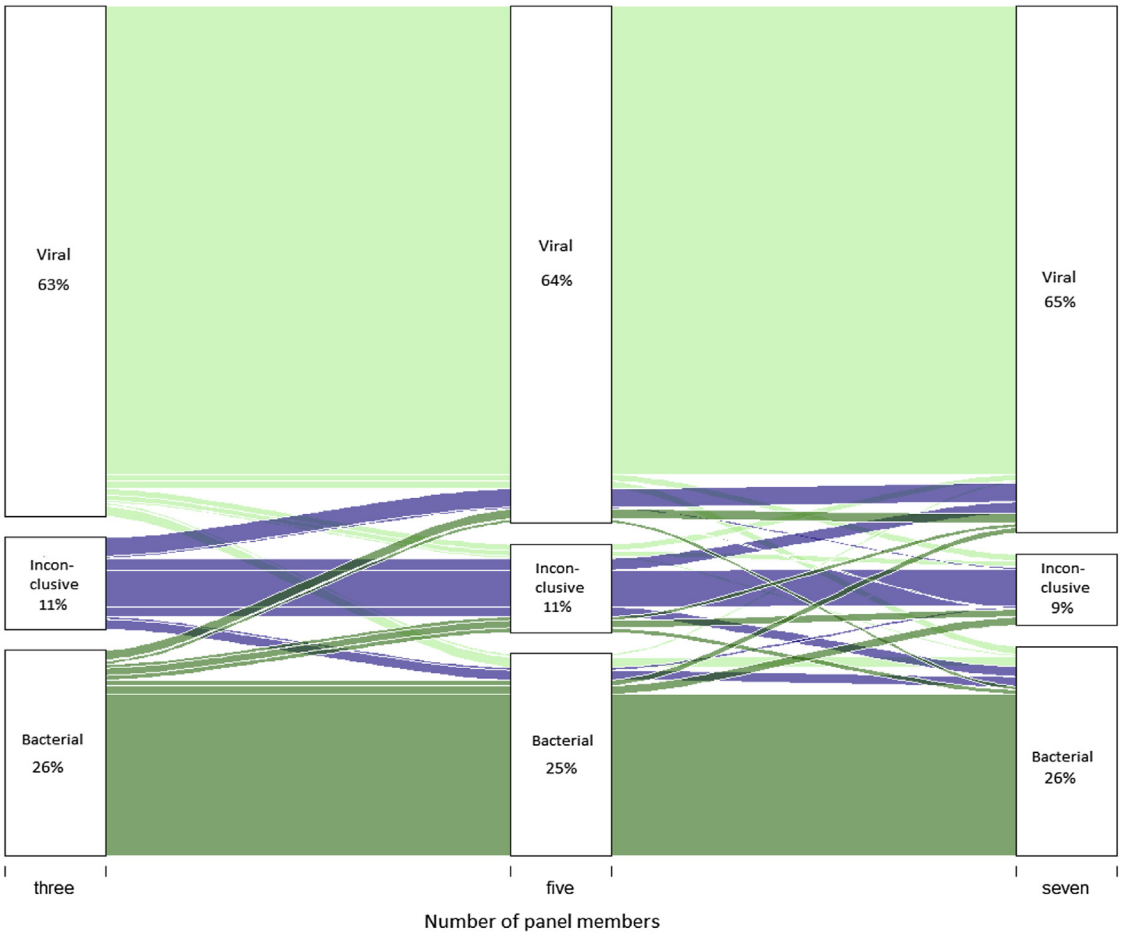

Fig. 3. Expert panel diagnosis using panels of different size. The waves show the change in diagnoses if the panel was expanded from three to five and seven experts (e.g., panel size analysis). Light green waves: cases with a viral diagnosis by a panel existing of three experts; purple waves: cases with an inconclusive diagnosis by a panel existing of three experts; dark green waves: cases with a bacterial diagnosis by a panel existing of three experts. Overall, diagnosis changed in 11% of the individual cases when the size of the panel increased from three to five experts and in 10% when expanding the panel size from five to seven experts. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

In the present study, a panel consisting of three experts as a reference standard in a diagnostic infectious diseases study provided more reproducible diagnoses than individual expert's diagnoses and showed a high level of intrapanel agreement. Expanding the panel to five or seven experts led to a substantial change in diagnoses and did not lead to a decrease in the number of inconclusive diagnoses.

A review by Tuijn et al. evaluated the agreement between health care professionals and examined the effectiveness of planned interventions for improving reproducibility. They found overall a fair agreement ($\kappa = 0.31$) between health care professionals before implementing various interventions [19]. This agreement is in line with ours and other studies which found a poor to moderate interrater agreement [4−10]. Several radiological and histological studies evaluated the intraobserver agreement. Most of these studies found high level of agreement [20−24]. However, it should be noted that these studies used relative short time intervals (i.e., several weeks) to assess intraobserver agreement. In our study, intraobserver agreement was lower for the 3 years' time interval compared to the 6 years' time interval. This result shows that it is crucial to mention the time interval when presenting intraobserver agreements. None of the previously mentioned studies evaluated the reproducibility of individual expert diagnosis compared to a panel. We found only two, noninfectious diseases-related studies that compared single observer diagnoses with panel diagnoses [15,16]. Both studies concluded, similarly to the findings in our study, that the reproducibility and accuracy of the diagnosis was higher for panel diagnosis and recommended this approach at least for research purposes.

Our findings may be valuable for researchers but also for organizations as the Food and Drug Administration (FDA) and European Medicines Agency (EMA), as these organizations approve drugs related to outcomes based on studies using single expert assessments [25,26]. However, our results demonstrated that if a panel diagnosis was used instead of individual expert diagnosis, the outcome of a study (i.e., the effect of the drug) might have changed significantly.

The major strength of our study is the comprehensive procedure of assessing reproducibility of a panel diagnosis in infectious diseases. To our knowledge, this is the first study that evaluated intraobserver agreement with such a long time interval (i.e., three years) and that examined the effect of expanding the size of the panel from three to seven experts. A second strength worth mentioning is the prospective character of the OPPORTUNITY study, from which cases were derived, leading to complete, longitudinal and high-quality data for every case.

Limitations of our study should also be discussed. First, we cannot exclude that experts included in the intraobserver and intrapanel analyses recognized cases from their first review. Obviously, the shorter the time interval, the more likely it is to recognize cases and to remember the previously assigned diagnosis. Although we have changed the study numbers, this may have contributed to the higher intraobserver and intrapanel agreement with the 6 weeks interval compared to the 3 years' time interval. Second, we presented an arbitrary set of data to the experts. We used the same eCRF as in our, previously published, validation study of a new diagnostic. Third, there are several aspects of panel diagnosis we did not evaluate in the present study, for example, the correlations between expert characteristics (i.e., experience years and country of origin) or eCRF variables (including specific biomarkers) and outcome. Fourth, the sample-size estimation was based on a convenient sample size. The two previously mentioned studies that compared single observer diagnoses with panel diagnoses used less cases compared to our study (i.e., 45 cases with dementia and 100 cases with seizures) [15,16]. Therefore, we feel our sample size is acceptable to obtain an adequate power to calculate expert panel reproducibility. Fifth, the review process of the cases by the experts required a significant effort from the expert panel members. Therefore, it was not feasible to have all 193 cases reviewed by the same seven experts for the second objective. This nonuniformity in the cases experts classified possibly impacted the results. Sixth, in our study, individual diagnosis changed in 11% of the cases when the size of the panel increased from three to five experts and in 10% when expanding the panel size from five to seven experts. We consider this as acceptable, as the proportion of inconclusive cases did not change. However, these differences might still have significant impact on the sensitivity and specificity of the panel diagnosis. Seventh, not all eligible patients participated in this study for practical reasons (e.g., attending physician did not have time to recruit the patient at the ED, parents or patients did not want phlebotomy only for study proposes), which may have introduced a selection bias in favor of more severely ill patients. Finally, given the wide variety of consensus panels, our findings cannot be generalized to noninfectious patients.

In conclusion, this study suggests that a panel consisting of three experts provides more reproducible diagnosis than an individual expert, while increasing the size of experts' panel further has no major advantage for diagnosis reproducibility. The impact of these results on validity could not be evaluated as a gold standard to diagnose bacterial infections is lacking. Based on our results, a panel diagnosis of experts, instead of the diagnosis of a single expert, is highly preferred as a reference standard. Further prospective validation and optimization of panel diagnosis are needed to obtain accurate results of diagnostic studies without a gold standard.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.03.010.

References

[1] Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009;62:797−806.

[2] Lynch T, Bialy L, Kellner JD, Osmond MH, Klassen TP, Durec T, et al. A systematic review on the diagnosis of pediatric bacterial pneumonia: when gold is bronze. PLoS One 2010;5:e11989.

[3] Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11. iii, ix-51.

[4] DiGiorgio MJ, Vinski J, Bertin M, Sun Z, Bena JF, Albert NM. Single-center study of interrater agreement in the identification of central line-associated bloodstream infection. Am J Infect Control 2014; 42(6):638−42.

[5] McBryde ES, Brett J, Russo PL, Worth LJ, Bull AL, Richards MJ. Validation of statewide surveillance system data on central line-associated bloodstream infection in intensive care units in Australia. Infect Control Hosp Epidemiol 2009;30(11):1045−9.

[6] Bada C, Carreazo NY, Chalco JP, Huicho L. Inter-observer agreement in interpreting chest X-rays on children with acute lower respiratory tract infections and concurrent wheezing. Sao Paulo Med J 2007; 125(3):150−4.

[7] Fischer JE, Seifarth FG, Baenziger O, Fanconi S, Nadal D. Hindsight judgement on ambiguous episodes of suspected infection in critically ill children: poor consensus amongst experts? Eur J Pediatr 2003; 162(12):840−3.

[8] Greenwald PW, Schaible DD, Ruzich JV, Prince SJ, Birnbaum AJ, Bijur PE. Is single observer identification of wound infection a reliable endpoint? J Emerg Med 2002;23(4):333−5.

[9] Loeb MB, Carusone SB, Marrie TJ, Brazil K, Krueger P, Lohfeld L, et al. Interobserver reliability of radiologists' interpretations of mobile chest radiographs for nursing home-acquired pneumonia. J Am Med Dir Assoc 2006;7(7):416−9.

[10] Klompas M. Interobserver variability in ventilator-associated pneumonia surveillance. Am J Infect Control 2010;38(3):237−9.

[11] de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. BMJ 2013;346:f2125.

[12] Gauvin F, Dassa C, Chaibou M, Proulx F, Farrell CA, Lacroix J. Ventilator-associated pneumonia in intubated children: comparison of different diagnostic methods. Pediatr Crit Care Med 2003;4(4):437−43.

[13] Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med 2013;10(10):e1001531.

[14] van Houten CB, de Groot JA, Klein A, Srugo I, Chistyakov I, de Waal W, et al. A host-protein based assay to differentiate between bacterial and viral infections in preschool children (OPPORTUNITY): a double-blind, multicentre, validation study. Lancet Infect Dis 2017;17:431−40.

[15] Stroink H, van Donselaar CA, Geerts AT, Peters AC, Brouwer OF, van Nieuwenhuizen O, et al. Interrater agreement of the diagnosis and classification of a first seizure in childhood. The Dutch Study of Epilepsy in Childhood. J Neurol Neurosurg Psychiatry 2004; 75(2):241−5.

[16] Gabel MJ, Foster NL, Heidebrink JL, Higdon R, Aizenstein HJ, Arnold SE, et al. Validation of consensus panel diagnosis in dementia. Arch Neurol 2010;67(12):1506−12.

[17] Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8(1): 23−34.

[18] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.

[19] Tuijn S, Janssens F, Robben P, van den Bergh H. Reducing interrater variability and improving health care: a meta-analytical review. J Eval Clin Pract 2012;18:887−95.

[20] Muhlhofer HM, Lenze U, Lenze F, Rondak IC, Schauwecker J, Rechl H, et al. Inter- and intra-observer variability in biopsy of bone and soft tissue sarcomas. Anticancer Res 2015;35(2):961−6.

[21] van Buijtenen JM, van Tunen ML, Zuidema WP, Heilbron EA, de Haan J, de Vet HC, et al. Inter- and intra-observer agreement of the AO classification for operatively treated distal radius fractures. Strategies Trauma Limb Reconstr 2015;10(3):155−9.

[22] Imerci A, Aydogan NH, Tosun K. Evaluation of inter- and intra-observer reliability of current classification systems for subtrochanteric femoral fractures. Eur J Orthop Surg Traumatol 2018;28(3): 499−502.

[23] Okizaki A, Nakayama M, Nakajima K, Katayama T, Uno T, Morikawa F, et al. Inter- and intra-observer reproducibility of quantitative analysis for FP-CIT SPECT in patients with DLB. Ann Nucl Med 2017;31(10):758−63.

[24] Ominde M, Sande J, Ooko M, Bottomley C, Benamore R, Park K, et al. Reliability and validity of the World Health Organization reading standards for paediatric chest radiographs used in the field in an impact study of Pneumococcal Conjugate Vaccine in Kilifi, Kenya. PLoS One 2018;13:e0200715.

[25] Meltzer HY, Risinger R, Nasrallah HA, Du Y, Zummo J, Corey L, et al. A randomized, double-blind, placebo-controlled trial of aripiprazole lauroxil in acute exacerbation of schizophrenia. J Clin Psychiatry 2015;76(8):1085−90.

[26] Citrome L, Du Y, Risinger R, Stankovic S, Claxton A, Zummo J, et al. Effect of aripiprazole lauroxil on agitation and hostility in patients with schizophrenia. Int Clin Psychopharmacol 2016;31(2): 69−75.