Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the Emergency Department.

Objectives To measure inter-rater agreement of overall clinical appearance of febrile children aged less than 24 months and to compare methods for doing so.

Study Design and setting We performed an observational study of inter-rater reliability of the assessment of febrile children in a county hospital emergency department serving a mixed urban and rural population. Two emergency medicine healthcare providers independently evaluated the overall clinical appearance of children less than 24 months of age who had presented for fever. They recorded the initial 'gestalt' assessment of whether or not the child was ill appearing or if they were unsure. They then repeated this assessment after examining the child. Each rater was blinded to the other's assessment. Our primary analysis was graphical. We also calculated Cohen's κ, Gwet's agreement coefficient and other measures of agreement and weighted variants of these. We examined the effect of time between exams and patient and provider characteristics on inter-rater agreement.

Results We analyzed 159 of the 173 patients enrolled. Median age was 9.5 months (lower and upper quartiles 4.9-14.6), 99/159 (62%) were boys and 22/159 (14%) were admitted. Overall 118/159 (74%) and 119/159 (75%) were classified as well appearing on initial 'gestalt' impression by both examiners. Summary statistics varied from 0.223 for weighted κ to 0.635 for Gwet's AC2. Inter rater agreement was affected by the time interval between the evaluations and the age of the child but not by the experience levels of the rater pairs. Classifications of 'not ill appearing' were more reliable than others.

Conclusion The inter-rater reliability of emergency providers' assessment of overall clinical appearance was adequate when described graphically and by Gwet's AC. Different summary

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

statistics yield different results for the same dataset.

Title: Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the Emergency Department.

Authors
Paul Walsh (1), Justin Thornton (2), Julie Asato (2), Gary McCoy (2), Nicholas Walker (2), Joe Baal (2), Jed Baal (2), Nanse Mendoza(2) Faried Banimahd (3)

Corresponding author
Paul Walsh
pfwalsh@ucdavis.edu

Affiliations
(1) Department of Emergency Medicine, University of California Davis Medical Center, 4150 V Street, Sacramento, CA 95817
(2) Department of Emergency Medicine, Kern Medical Center, 1830 Flower Street, Bakersfield, CA 93305.
(3) Department of Emergency Medicine, University of California Irvine, Orange, CA

Keywords
Inter-rater agreement, Gwet's AC, Cohen's kappa, Graphical analysis, Pediatric, Emergency medicine, Fever

Title: Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the Emergency Department.

Abstract

Objectives

To measure inter-rater agreement of overall clinical appearance of febrile children aged less than 24 months and to compare methods for doing so.

Study Design and setting

We performed an observational study of inter-rater reliability of the assessment of febrile children in a county hospital emergency department serving a mixed urban and rural population. Two emergency medicine healthcare providers independently evaluated the overall clinical appearance of children less than 24 months of age who had presented for fever. They recorded the initial 'gestalt' assessment of whether or not the child was ill appearing or if they were unsure. They then repeated this assessment after examining the child. Each rater was blinded to the other's assessment. Our primary analysis was graphical. We also calculated Cohen's κ, Gwet's agreement coefficient and other measures of agreement and weighted variants of these. We examined the effect of time between exams and patient and provider characteristics on inter-rater agreement.

Results

We analyzed 159 of the 173 patients enrolled. Median age was 9.5 months (lower and upper quartiles 4.9-14.6), 99/159 (62%) were boys and 22/159 (14%) were admitted. Overall 118/159 (74%) and 119/159 (75%) were classified as well appearing on initial 'gestalt' impression by both examiners. Summary statistics varied from 0.223 for weighted κ to 0.635 for Gwet's AC2. Inter rater agreement was affected by the time interval between the evaluations and the age of the child but not by the experience levels of the rater pairs. Classifications of 'not ill appearing' were more reliable than others.

Conclusion

The inter-rater reliability of emergency providers' assessment of overall clinical appearance was adequate when described graphically and by Gwet's AC. Different summary statistics yield different results for the same dataset.

INTRODUCTION

Deciding whether a febrile child is 'ill appearing' is a key decision point in emergency department (ED) management algorithms for febrile infants and toddlers.(Baker et al. 1993, Baraff et al. 1993, Jaskiewicz et al. 1994, Baskin et al. 1992) Initial physician judgments of this overall appearance are generally made rapidly and prior to completing a full physical examination. Such judgments can even affect how providers interpret clinical findings.(McCarthy et al. 1985)

Implicit in this construct is the assumption that clinicians agree on whether or not a child is ill appearing. There is little evidence that addresses the inter-rater reliability of providers' overall impression of febrile children's appearance. One study found good agreement for individual clinical signs, many of which are associated with overall clinical appearance and often with fever.(Wagai et al. 2009) Others have addressed inter-rater reliability for the Yale observation score(McCarthy et al. 1985); but studies of overall clinical impression without the use of specific scoring systems are scarce. The inter-rater reliability of individual historical and examination findings has been studied for a variety of conditions including diagnostic interviews, head trauma and bronchiolitis.(Shaffer et al. 1993) (Holmes et al. 2005) (Walsh et al. 2006) Establishing adequate inter-rater reliability is an important part in the derivation of clinical management algorithms (Laupacis A 1997, Stiell and Wells 1999) but is often not performed. (Maguire et al. 2011)

Although clinical appearance is a binary decision node in management algorithms,(Baker et al. 1993, Baraff et al. 1993) (Jaskiewicz et al. 1994) clinical appearance is a continuum as some children appear more ill than others. When given the option providers chose 'unsure' in 12.6% of infants and toddlers presenting to an ED in one study.(Walsh et al. 2014) These children in whom the provider was "unsure" had historical and physical examination findings intermediate in severity between those classified as ill and not ill appearing. The prevalence of bacterial meningitis and pneumonia was also intermediate between those classified as ill or not ill appearing.(Walsh et al. 2014)

Despite the widespread use of management strategies that rely on overall clinical appearance, the inter-rater reliability of clinical appearance is not well established. Moreover, because ill appearing children are in a small minority, widely used measures of inter-rater reliability such as Cohen's κ statistic risk being overly conservative. This is also true for other summary measures of inter rater agreement which rely on the marginal distribution of categories. Consequently even though actual agreement (reliability) between raters is high the summary statistic will be low. In the context of clinical decision making this could lead to useful clinical characteristics being incorrectly labeled too unreliable for clinical use. Alternative approaches, including simple graphical analysis exist but are not widely used in emergency medicine.

The first aim of this study was to measure inter-rater agreement of overall clinical appearance of febrile children aged less than 24 months. We hypothesized that inter-rater agreement of overall clinical appearance would be adequate for clinical use. In addition, we hypothesized that agreement is influenced by the clinical experience of raters. The second aim of this study was to compare methods for evaluating inter-rater agreement in unbalanced samples and in particular examine graphical methods.

METHODS

Design

This was a cross sectional, prospective observational study of inter-rater reliability performed in accordance with the guidelines for reporting reliability and agreement studies. (Kottner et al. 2011) The study was approved by Kern Medical Center institutional review board (approval #10032). We obtained verbal consent and provided informational materials in lieu of written consent from parents or legal guardians and providers.

Setting

The study was performed at a county hospital teaching emergency department (ED) with emergency medicine residency.

Subjects

The subjects were 9 board eligible or certified general emergency medicine physicians, a pediatric emergency physician, three mid-level providers and 21 emergency medicine residents for a total of 34 providers. The patients in the study were children aged less than 24 months who presented to the ED with a chief complaint of fever, or a rectal temperature of at least 38ºC at triage.

Implementation

Eligible patients were identified by research assistants (RA) or physician investigators. RA coverage was typically available 12-16 hours a day, seven days a week including at night and on holidays. Two physicians or mid-level providers were asked to give their assessment of infant appearance using the categories 'ill appearing', 'not ill appearing' or ' not sure'. Providers were asked to do this both before and again after examining the child. Each provider performed their evaluation without the other being present and blinded to the other provider's results. The data were recorded on two identical data forms, one for each provider. RAs entered the results into a customized database (Filemaker Pro, Santa Clara, CA).

Rationale for statistical methods

Simple percentage agreement may be an adequate measure of agreement for many purposes, but does not account for agreement arising from chance alone. Attempts to account for the agreement that may arise from chance have led to a variety of methods for measuring inter-rater reliability. These methods vary with different approaches for continuous, ordinal, categorical, and nominal data (Cohen 1960, Fleiss 1971, Cohen 1968, Banerjee et al. 1999). Our data could be considered categorical but, based on the ordinal association between components of clinical appearance and some microbiological outcomes, our classification scheme could also be considered ordinal. We used both approaches.

Categorical agreement is often measured with Cohen's κ. Cohen's κ appears easily interpreted; its minimum and maximum are -1 and +1 for 2x2 tables. For a k x k table the minimum is -1/(k-1) and approaches 0 from the bottom as k gets larger; while the maximum is always +1. Negative values imply disagreement beyond independence of raters and positive values agreement beyond independence of raters. Descriptive terms such as 'moderate' and 'poor' agreement have been published to further ease interpretation.(Landis and Koch 1977) The simple κ assumes two unique raters. When the two raters' identities vary an implementation of the more than two raters case must be used. (Statacorp 2013, Fleiss et al. 2003) A provider could be the first reviewer for one infant and the second reviewer for another. We did this because our question was about the inter-rater reliability of providers in general rather than any specific provider pair. Consequently the study we carried out was one of many that could have been carried out; by reversing the order of which provider was selected as reviewer 1 and reviewer 2 one could conceivably obtain different κ scores even though the percentage agreement would be unchanged. Assuming there was no bias in how we selected first and second reviewers we anticipated this effect would be small given the kappa calculation we used. We simulated 500 alternative permutations of the order of reviewers to verify this assumption.

The best design for a k x k table to measure agreement is one where the margins have roughly a proportion of 1/k of the total sample studied; in the 2x2 case this means a prevalence of 0.5 for the indication as well as its compliment. Serious deviations from this are known to make the variance for κ unstable and κ misleading for measuring agreement amongst the k levels of the scale. However such samples are unrepresentative of most clinical scenarios, particularly in emergency medicine where non-serious outcomes often far outnumber serious ones. A disadvantage of the κ statistic is that it results in lower values the further the prevalence of the outcome being studied deviates from 0.5.(Feinstein and Cicchetti 1990, Gwet 2008) Scott's π (subsequently extended by Fleiss) suffers the same limitations.(Scott 1955) This so called 'κ paradox' is well described and understood by statisticians. When interpreted by others however, this property of κ could lead to clinical tools with potentially useful but imperfect reliability being discarded based on a low reported κ value. Consequently κ and Scott's π risk misinterpretation when one of the categories being rated is much more or less common than the other. Gwet developed an alternative method, the agreement coefficient (AC_1) specifically to address this limitation.(Gwet 2008) The AC_1 has potential minimum and maximum values of -1 and +1 respectively. The AC_1 is more stable than κ although the AC_1 may give slightly lower estimates than κ when the prevalence of a classification approaches 0.5 but gives higher values otherwise. (Gwet 2008) The AC_1 does not appear widely in the medical literature despite recommendations to use it.(Wongpakaran et al. 2013, McCray 2013) This may be because of two key assumptions, namely that chance agreement occurs when at least one rater rates at least some individuals randomly and that the portion of the observed ratings subject to randomness is unknown. On the other hand these assumptions may not be stronger than those inherent in Cohen's κ.

Ordinal agreement can be measured using a weighted κ. The penalty for disagreement is weighted according to the number of categories by which the raters disagree.(Cohen 1968) The results are dependent both on the weighting scheme chosen by the analyst and the relative prevalence of the categories.(Gwet 2008) One commonly recommend weighting scheme reduces the weighted κ to an intra-class correlation.(Fleiss and Cohen 1973) Scott's π and Gwet's AC_1 can also be weighted. When weighted, Gwet's AC_1 is referred to as AC_2.(Gwet 2012)

Another approach is to regard ordinal categories as bins on a continuous scale. Polychoric correlation estimates the correlation between raters as if they were rating on a continuous scale. (Flora and Curran 2004, Uebersax 2006) Polychoric correlation is at least in principle insensitive to the number of categories and can even be used where raters use different numbers of categories. The correlation coefficient, -1 to +1, is interpreted in the usual manner. A disadvantage of polychoric correlation is that it is susceptible to distribution; although some recognize polychoric correlation as a special case of latent trait modeling thereby allowing relaxation of distribution assumptions. (Uebersax 2006) The arguments against using simple correlation as a measure for agreement for continuous variables in particular have been well described.(Bland and Altman 1986)

It is easy to conceive that well appearing infants are more common than ill appearing ones, thereby raising concerns that assumptions of a normal distribution are unlikely to hold. Another coefficient of agreement "A" proposed by van der Eijk was specifically designed for ordinal scales with a relatively small number of categories dealing with abstract concepts. This measure "A" is insensitive to standard deviation. "A" however contemplates large numbers of raters rating a small number of subjects (such as voters rating political parties).(Van der Eijk 2001)

The decision to use of a single summary statistic to describe agreement is fraught with the risks of imbalance in the categories being rated, different results from different methods and the need to ordain in advance a specific threshold below which the characteristic being classified will be discarded as too unreliable to be useful for decision making.

We used a simple graphical method for our primary analysis. For the graphical method we categorized agreement as follows:

Reviewer 1 and reviewer 2 agree	Ill appearing : ill appearing Not ill appearing : not ill appearing Unsure : unsure
Reviewer 1 considers infant more ill appearing by one category than reviewer 2	Ill appearing : unsure Unsure : Not ill appearing
Reviewer 1 considers patient more ill appearing by two categories than reviewer 2	Ill appearing : Not ill appearing
Reviewer 1 considers patient less ill appearing by one category than reviewer 2	Unsure : ill appearing Not ill appearing : unsure
Reviewer 1 considers patient more ill appearing by two categories than reviewer 2	Not ill appearing : Ill appearing

We created a bar graph with a bar representing the percentage of patients in each category and, by simulation, (bottom right in the figure) a graph to portray how random assignment of categories would appear. This graph would be expected to be symmetrical around the bar portraying when the providers agreed. Asymmetry could suggest bias or suggest or that a change in the quantity being measured has occurred between the two exams. This could arise if the infants' condition changed between the two exams. We created an artificial dataset where agreement was uniformly randomly assigned and used this to create a reference graph of what random agreement alone would look like. All graphs were drawn using Stata 13.

We also calculated weighted kappa (κ) using widely used weighting schemes, polychoric correlation, and Gwet's agreement coefficient (AC_1 and AC_2) as secondary methods.(Gwet 2008) We performed sensitivity analysis using logistic regression to examine the effect age, diagnosis, antipyretic use, experience levels of the raters, and time between evaluations. We analyzed the rater pairs as using several strategies. In one we assigned an interval value for each year of post graduate training with attending physicians all assigned a value of six. In another strategy we grouped residents as PGY1 and PGY2, PGY3 and PGY4, MLP and attending, assigned values of 1 to 4 and analyzed these. We also examined rater pair combinations as nominal variables.

Sample Size Calculations

Given the lack of sample size methods for graphical analysis we relied on sample size calculations for a traditional Cohen's κ. We assumed that 75% of each raters' classifications would be for the more common outcome, a κ of 0.8, an absolute difference range in the final κ of +/- 0.1 and an α of 0.05.(Reichenheim 2000) This resulted in a sample size of 144 patients.

Data management, logistic, κ and polychoric(Kolenikov 2004) estimations were performed using Stata version 13.0 software (Statacorp LLP, College Station, TX). Gwet's AC1 was calculated using R Version 3.01, (www.r-project.org. AC1 function from Emmanuel).(Emmanuel 2013) Other measures of agreement were estimated using Agreestat, (www.agreestat.com).

RESULTS

We analyzed 159 of the 173 patients enrolled. Patient flow and reasons for patient exclusion are shown in **Figure 1**. There were 99/159 (62%) boys and the median age was 9.5 months, (lower and upper quartiles 4.9, 14.6 months) and 22/159 (14%) were admitted. Eighty (50%) patients received antipyretics prior to evaluation by both providers, and 25/159 (16%) received antipyretics between the first and second provider's assessments. The ED diagnoses are summarized in **Table 1**.

We observed 29 different combinations in the order of evaluations and level of provider training. These are described using a density distribution sunflower plot(Dupont and Plummer 2003) in **Figure 2**. Overall 118/159 (74%) and 119/159 (75%) were classified as well appearing on initial 'gestalt' impression by the two examiners. When the first rater classified a child as well appearing the second was more likely to agree (94/120) (78%) than when the first rater classified the child as ill appearing 8/27 (30%) p=0.025. The agreement between all raters and categories are shown in **Table 2** and **Table 3;** intra rater agreement is shown in **Table 4**.

The weighted κ was 0.223 for initial gestalt assessment and 0.314 following examination. Our simulation comparing 500 random alternative permutations of first and second reviewers found little evidence of bias. Where our observed weighted κ was 0.223 the minimum and maximum found in our alternative possible reviewer order permutations was 0.220 to 0.235. This argues against bias in the order in which the reviewers were selected.

The polychoric correlation for initial 'gestalt' assessment and assessment following examination were 0.334 and 0.482 respectively. These, the weighted κ, and the AC_2 all point to increased agreement after the examiners had completed a full exam of the infant. When doctors differ in their 'gestalt' evaluation of a febrile child's overall appearance both of them doing a detailed examination of the patient will narrow their differences.

The frequency with which providers of different training levels chose each classification is shown in **Figure 3**. However despite none of our analyses demonstrated a significant effect for level of training and agreement. Inter and intra-rater agreement is shown in **Figure 4.** Inter-rater agreement improved with examination compared to gestalt assessment. **Table 5** (further expanded in **Appendix 3** provides various κ, π, polychoric and AC_1 and AC_2 statistics for the results in **Tables 2-4.** All of these point to increasing agreement when more clinical information is obtained. This also suggests a practical solution for clinicians when faced with uncertainty, either go back and examine the child again or ask a colleague to do so.

There was some asymmetry in the graphs portraying intra rater reliability (**Figure 4**). This suggests that a full exam may lead a provider to revise their impression toward increasing the severity of the child's appearance of illness more often than revising their impression towards decreasing the severity of the child's appearance of illness.

There was also slight asymmetry in the graphs describing inter rater reliability favoring increase in the overall appearance of illness by the second reviewer. Sensitivity analysis showed that age (months), odds ratio (OR) 0.90, (95%CI 0.84, 0.95) and time between evaluations (10 minute intervals), OR 0.92 (95% CI 0.85, 0.99) impacted inter-rater agreement. Antipyretic use (even when interacted with the time interval between evaluations), experience of the provider pair, diagnosis, and infant age less than 2 months, were all non-significant. The inter-rater agreement of overall clinical impression between providers was greatest when the child was considered 'not ill appearing'.

We found a wide range of values that could be calculated for different κ variants and other measures of agreement and the very low values of traditional marginal agreement statistics. Many of these could reasonably be presented as a true reflection of the inter rater-reliability of provider's assessment of a febrile child's overall appearance. Based on most of the inter rater reliability measures a central tenant of the management of febrile infants and toddlers would be discarded as unreliable. However our graphical analysis portrays a different picture entirely. This picture is one of overwhelming agreement with the caution that a second or closer look may find evidence of increasingly ill appearance. Of the summary statistics of agreement only Gwet's AC provided an estimate that would allow a reader to intuit the agreement observed given the observed imbalance between ill and not ill appearing children.

DISCUSSION

The inter-rater reliability of ED provider assessment of overall clinical appearance in febrile children aged less than 24 months was modest. Inter-rater agreement was better after providers had examined the child than when they relied solely on gestalt. Agreement decreased in older children, and as the time interval between the two providers' evaluations increased. Classifications of 'not ill appearing' were more reliable than others. Provider experience level had little effect on agreement.

Different summary measures of agreement, and different weighting schemes yielded different results. Graphical portrayal of these results better communicated the inter-rater reliability than did the single summary statistical measures of agreement. Among the summary statistical measures of agreement Gwet's AC most closely paralleled the graphical presentation of results.

These results are broadly consistent with those of Wagai *et al* who compared clinicians' evaluations using videos of children.(Wagai et al. 2009) We have previously used videos for training and measuring inter-rater reliability in the Face, Legs, Activity, Cry and Consolability (FLACC) pain score in infants. Videos allow standardization of inter-rater reliability measurements. The disadvantage of videos, however, is a loss of validity as an artificial situation is created. Our finding that clinical experience did not affect agreement of overall clinical appearance is consistent with the finding of Van der Bruel who found that the seniority of the physician did not affect the diagnostic importance of a 'gut feeling that something is wrong' in 3,981 patients aged 0-16 years.(Van den Bruel et al. 2012) Our findings differ from prior work in children aged less than 18 months with bronchiolitis.(Walsh et al. 2006) This may be because of inherent differences in the conditions.

The use of the κ statistic is often an appropriate strategy for analyzing studies of inter-rater reliability. However the apparent paradox where high actual agreement can be associated with a low or even negative κ can mislead rather than enlighten. This was clearly evident in our study, the weighted κ was 0.223. Experiments where examiners have been told to guess their physical findings based on a third party clinical report have been used to argue that low κ statistics in fact reflect the true reliability of examination findings in children.(Gorelick and Yen 2006) The difficulty for clinicians accepting such a strategy is that it appears to lack face validity.

Clinical outcomes do not demonstrate the variability in outcome that one would intuitively expect were clinical examination so unreliable.

We have also shown how differently weighted and κ and other measures of agreement may differ substantially from each other. (Gwet 2012) Some have recommended focusing on those cases in which disagreement might be anticipated,(Lantz and Nebenzahl 1996) (Vach 2005) others recommend abandoning κ (or the proposed experiment) entirely where expected prevalence is not ~50%.(Hoehler 2000) A middle ground approach has been to argue that in addition to the omnibus κ, percentages of positive and negative agreement should also be presented. (Cicchetti and Feinstein 1990)

This approach is not dissimilar to our graphical solution; a simple graph is readily grasped and allows the reader detect asymmetry which may suggest changes between the ratings. Portraying simple differences in agreement graphically may not however be the optimal solution for every situation. Graphs have the attraction of allowing relative complex data be absorbed quickly by readers, even if the reader has little or no statistical training.

Graphical methods also have disadvantages. Graphs require more space than a single summary statistic and are more difficult to summarize. A number is easier to communicate verbally to a colleague than a graph. Different readers may view the same graph and disagree about its meaning raising the question by whose eye should a graph be judged? Another limitation of graphical methods is their vulnerability to axis manipulation or failure to include a reference graph of what agreement by random chance alone would look like using the same scale for each axis.

The apparent objectivity and simplicity of a single number makes decision making easier. However we argue that summary statistics also increases the risk of the wrong decision being made as to whether or not a characteristic is sufficiently reliable be included in decision making. When graphical methods are not optimal, providing separate summaries of the proportionate agreement in each class, or Gwet's $AC_{1\&2}$ may be an alternative. Certainly, it seems unwise to discount clinical findings for inclusion in prediction rules and management algorithms solely based on κ scores of < 0.5, without consideration of other measures This is particularly the case where categories are expected to be highly unbalanced, as in for example serious bacterial infection in infants with bronchiolitis, intracranial bleeding in head injury and cervical spine injury in blunt trauma.(Leonard et al. 2011, Kuppermann et al. 2009, Chee et al. 2010)

Limitations

There are several limitations to our work. Data were collected at a single teaching hospital and the results may not be generalizable to other sites. Although a diverse mix of providers was measured, the lack of attending physician to attending physician comparisons decreases generalizability. We also assumed that within categories of raters are interchangeable. This is assumption is typical of research evaluating the reliability of clinical signs for inclusion in diagnostic or treatment algorithms. Few very sick infants were included. Similarly for the 'not sure' category our number of patients was very small. This may be because of the rarity of conditions such as bacterial meningitis and, perhaps because physicians' unwillingness to enroll very sick infants in the study out of concern that it would delay care or disposition. Such concerns have hampered previous attempts at measuring the inter rater reliability of the Ottawa ankle rule in children and may help explain the dearth of inter rater studies in the evaluation of febrile infants.(Plint et al. 1999)

CONCLUSION

The inter-rater reliability of EP assessment of overall clinical appearance was adequate. Inter-rater reliability is sometimes better described graphically than by a summary statistic; different summary statistics yield different results for the same dataset. Inter-rater agreement of overall appearance should not always be reduced to a single summary statistic but when categories are unbalanced Gwet's AC is preferred.

REFERENCES

Baker, M. D., Bell, L. M. and Avner, J. R. (1993) 'Outpatient management without antibiotics of fever in selected infants', *New England Journal of Medicine,* 329(20), 1437-1441.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999) 'Beyond kappa: A review of interrater agreement measures', *Canadian Journal of Statistics,* 27(1), 3-23.

Baraff, L. J., Schriger, D. L., Bass, J. W., Fleisher, G. R., Klein, J. O., McCracken, G. H. and Powell, K. R. (1993) 'Practice guideline for the management of infants and children 0 to 36 months of age with fever without sSource', *Pediatrics,* 92(1), 1-12.

Baskin, M. N., O'Rourke, E. J. and Fleisher, G. R. (1992) 'Outpatient treatment of febrile infants 28 to 89 days of age with intramuscular administration of ceftriaxone', *The Journal of Pediatrics,* 120(1), 22-27.

Bland, M.J. and Altman, D. (1986) 'Statistical methods for assessing agreement between two methods of clinical assessment', *The Lancet,* 327(8476), 307-310.

Chee, C., Walsh, P., Kuan, S., Cabangangan, J., Azimian, K., Dong, C., Tobias, J. and Rothenberg, S. J. (2010) 'Emergency department septic screening in respiratory Syncytial virus (RSV) and non-RSV bronchiolitis', *Western Journal of Emergency Medicine,* 11(1), 60-67.

Cicchetti, D. V. and Feinstein, A. R. (1990) 'High agreement but low kappa: II. Resolving the paradoxes', *Journal of Clinical Epidemiology,* 43(6), 551-558.

Cohen, J. (1960) 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement,* 20(1), 37-46.

Cohen, J. (1968) 'Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit', *Psychological Bulletin,* 70(4), 213-220.

Dupont, W. and Plummer, W. (2003) 'Density distribution sunflower plots', *Journal of Statistical Software,* 8(3), 1-5.

Emmanuel , P. (2013) AC1 for R, statistical code. https://stat.ethz.ch/pipermail/r-sig-epi/attachments/20120503/7c03297b/attachment.pl [accessed 7/16/2014.]

Feinstein, A. R. and Cicchetti, D. V. (1990) 'High agreement but low kappa: I. The problems of two paradoxes', *Journal of Clinical Epidemiology,* 43(6), 543-549.

Fleiss, J., Levin, B. and Paik, M. (2003) *Statistical Methods for rates and proportions,* New York: Wiley.

Fleiss, J. L. (1971) 'Measuring nominal scale agreement among many raters', *Psychological bulletin,* 76(5), 378-382.

Fleiss, J. L. and Cohen, J. (1973) 'The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability', *Educational and psychological measurement*.

Flora, D. B. and Curran, P. J. (2004) 'An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data', *Psychological Methods,* 9(4), 466-491.

Gorelick, M. H. and Yen, K. (2006) 'The kappa statistic was representative of empirically observed inter-rater agreement for physical findings', *Journal of Clinical Epidemiology,* 59(8), 859-861.

Gwet, K. L. (2008) 'Computing inter-rater reliability and its variance in the presence of high agreement', *British Journal of Mathematical and Statistical Psychology,* 61(1), 29-48.

Gwet, K. L. (2012) *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters,* Advanced Analytics Press.

Hoehler, F. K. (2000) 'Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity', *Journal of Clinical Epidemiology,* 53(5), 499-503.

Holmes, J. F., Palchak, M. J., MacFarlane, T. and Kuppermann, N. (2005) 'Performance of the pediatric Glasgow coma scale in children with blunt head trauma', *Academic Emergency Medicine,* 12(9), 814-819.

Jaskiewicz, J. A., McCarthy, C. A., Richardson, A. C., White, K. C., Fisher, D. J., Powell, K. R. and Dagan, R. (1994) 'Febrile infants at low risk for serious bacterial infection—an appraisal of the Rochester criteria and implications for management', *Pediatrics,* 94(3), 390-396.

Kolenikov, S. (2004) polychoric : A Stata command to perform polychoric correlations, statistical code. http://www.unc.edu/~skolenik/stata/ [accessed 7/16/2014.]

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M. and Streiner, D. L. (2011) 'Guidelines for reporting reliability and agreement studies (GRRAS) were proposed', *International Journal of Nursing Studies,* 48(6), 661-671.

Kuppermann, N., Holmes, J. F., Dayan, P. S., Hoyle, J. D., Atabaki, S. M., Holubkov, R., Nadel, F. M., Monroe, D., Stanley, R. M., Borgialli, D. A., Badawy, M. K., Schunk, J. E., Quayle, K. S., Mahajan, P., Lichenstein, R., Lillis, K. A., Tunik, M. G., Jacobs, E. S., Callahan, J. M., Gorelick, M. H., Glass, T. F., Lee, L. K., Bachman, M. C., Cooper, A., Powell, E. C., Gerardi, M. J., Melville, K. A., Muizelaar, J. P., Wisner, D. H., Zuspan, S. J., Dean, J. M. and Wootton-Gorges, S. L. (2009) 'Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study', *The Lancet,* 374(9696), 1160-1170.

Landis, J. R. and Koch, G. G. (1977) 'The measurement of observer agreement for categorical data', *Biometrics*, 33(1),159-174.

Lantz, C. A. and Nebenzahl, E. (1996) 'Behavior and interpretation of the κ statistic: Resolution of the two paradoxes', *Journal of Clinical Epidemiology,* 49(4), 431-434.

Laupacis A, S. N. S. l. (1997) 'Clinical prediction rules: A review and suggested modifications of methodological standards', *JAMA,* 277(6), 488-494.

Leonard, J. C., Kuppermann, N., Olsen, C., Babcock-Cimpello, L., Brown, K., Mahajan, P., Adelgais, K. M., Anders, J., Borgialli, D., Donoghue, A., Hoyle Jr, J. D., Kim, E., Leonard, J. R., Lillis, K. A., Nigrovic, L. E., Powell, E. C., Rebella, G., Reeves, S. D., Rogers, A. J., Stankovic, C., Teshome, G. and Jaffe, D. M. (2011) 'Factors associated with cervical spine Injury in children after blunt trauma', *Annals of Emergency Medicine,* 58(2), 145-155.

Maguire, J. L., Kulik, D. M., Laupacis, A., Kuppermann, N., Uleryk, E. M. and Parkin, P. C. (2011) 'Clinical prediction rules for children: A systematic review', *Pediatrics,* 128(3), e666-e677.

McCarthy, P. L., Lembo, R. M., Baron, M. A., Fink, H. D. and Cicchetti, D. V. (1985) 'Predictive value of abnormal physical examination findings in ill-appearing and well-appearing febrile children', *Pediatrics,* 76(2), 167-171.

McCray, G. (2013) 'Assessing inter-rater agreement for nominal judgement variables.', in Nottingham, UK, University of Lancaster, Talk. Availbale at http://www.norbertschmitt.co.uk/uploads/27_528d02015a6da191320524.pdf , [accessed 7/17/14.]

Plint, A. C., Bulloch, B., Osmond, M. H., Stiell, I., Dunlap, H., Reed, M., Tenenbein, M. and Klassen, T. P. (1999) 'Validation of the Ottawa Ankle Rules in Children with Ankle Injuries', *Academic Emergency Medicine,* 6(10), 1005-1009.

Reichenheim, M. (2000) 'Sample size for the kappa-statistic of interrater agreement', *Stata Technical Bulletin* 58(1), 41-47.

Scott, W. A. (1955) 'Reliability of content analysis: The case of nominal scale coding', *Public opinion quarterly,19(3),321-325*.

Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Placentini, J., Davies, M., Conners, C. K. and Regier, D. (1993) 'The diagnostic interview schedule for children-Revised version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability', *Journal of the American Academy of Child and Adolescent Psychiatry,* 32(3), 643-650.

Statacorp (2013) *Stata 13 Base Reference Manual, Stata 13 Base Reference Manual* College Station, TX. : Stata Press.

Stiell, I. G. and Wells, G. A. (1999) 'Methodologic standards for the development of clinical decision rules in emergency medicine', *Annals of Emergency Medicine,* 33(4), 437-447.

Uebersax, J. S. (2006) ' The tetrachoric and polychoric correlation coefficients. Statistical methods for rater agreement. web site', [online], available: http://john-uebersax.com/stat/tetra.htm [accessed 7/16/2014].

Vach, W. (2005) 'The dependence of Cohen's kappa on the prevalence does not matter', *Journal of Clinical Epidemiology,* 58(7), 655-661.

Van den Bruel, A., Matthew, T., Frank, B. and David, M. (2012) 'Clinicians' gut feeling about serious infections in children: observational study', *BMJ,* 345,e6144.

Van der Eijk, C. (2001) 'Measuring agreement in ordered rating scales', *Quality and Quantity,* 35(3), 325-341.

Wagai, J., Senga, J., Fegan, G. and English, M. (2009) 'Examining agreement between clinicians when assessing sick children', *PLoS ONE,* 4(2), e4626.

Walsh , P., Capote, A., Garcha, D. and al, e. (2014) 'The Kern Fever Study: The meaning of incertitude when evaluating fever in the emergency department', in submission

Walsh, P., Gonzales, A., Satar, A. and Rothenberg, S. J. (2006) 'The interrater reliability of a validated bronchiolitis severity assessment tool', *Pediatric emergency care,* 22(5),316-320.

Wongpakaran, N., Wongpakaran, T., Wedding, D. and Gwet, K. L. (2013) 'A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples', *BMC medical research methodology,* 13(1), 61.

Table 1(on next page)

Diagnoses by classification

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

			First reviewer 'Gestalt' assessment				Second reviewer 'Gestalt' assessment		
Diagnosis			Ill appearing	Not Sure	Not ill appearing		Ill appearing	Not Sure	Not ill appearing
	n	**(%)**							
Pneumonia	14	(8.8)	3	3	8		3	1	10
UTI/Pyelonephritis	4	(2.5)	1	0	3		1	0	3
Bronchiolitis	9	(5.7)	2	0	7		2	1	6
Otitis media	6	(3.8)	1	1	4		1	1	4
Gastroenteritis	8	(5.0)	1	1	6		0	0	8
Cellulitis	6	(3.8)	4	0	2		3	0	3
Sepsis No focus	3	(1.9)	1	0	2		1	0	2
URI	36	(22.6)	3	3	30		5	1	30
Herpangina	2	(1.3)	0	0	2		0	0	2
Pharyngitis	2	(1.3)	0	0	2		0	0	2
Viral/Febrile illness NOS	46	(28.9)	5	5	36		4	5	37
Bacteremia	1	(0.6)	0	0	1		0	0	1
Varicella	2	(1.3)	0	1	1		0	0	2
Febrile Seizure	5	(3.1)	2	0	3		2	0	3
Non infective	3	(1.9)	0	0	3		0	0	3
Other Febrile illness	12	(7.6)	4	0	8		5	0	7
Total	**159**	**(100)**	27	14	118		27	9	123

Table 2 (on next page)

Inter rater reliability of ‘gestalt’ impression of overall clinical appearance

Inter rater reliability of ‘gestalt’ impression of overall clinical appearance with row and column percentages

First rater Gestalt Impression	Second rater Gestalt Impression							
	Not ill Appearing (%)		Not Sure (%)		Ill Appearing (%)		Total (%)	
Not ill Appearing	94 (80)	(78)	11 (9)	(69)	13 (11)	(57)	118 (100)	(74)
Not Sure	12 (86)	(10)	0 (0)	(0)	2 (14)	(9)	14 (100)	(9)
Ill Appearing	14 (52)	(12)	5 (19.5)	(31)	8 (30.5)	(35)	27 (100)	(17)
Total	120 (75.5)	(100)	16 (10)	(100)	23 (14.5)	(100)	159 (100)	(100)

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Table 3(on next page)

Inter rater agreement of overall clinical appearance after examining the patien

Inter rater agreement of overall clinical appearance after examining the patient with row and column percentages

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

First rater Impression after examining	Second rater Impression after examining											
	Not ill Appearing (%)			Not Sure (%)			Ill Appearing (%)			Total (%)		
Not Appearing	103	(84)	(82)	6	(5)	(75)	14	(11)	(54)	123	(100)	(77)
Not Sure	8	(89)	(6)	0	(0)	(0)	1	(11)	(4)	9	(100)	(6)
Ill Appearing	14	(52)	(11)	2	(7)	(25)	11	(41)	(42)	27	(100)	(17)
Total	125	(79)	(100)	8	(5)	(100)	26	(16)	(100)	159	(100)	(100)

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Table 4(on next page)

Intra-rater reliability for first (4A) and second raters (4B)

Intra-rater reliability for first (4A) and second raters (4B)

Table 4A

First rater Gestalt Impression	**First rater Impression after examining**							
	Not ill Appearing (%)		**Not Sure (%)**		**Ill Appearing (%)**		**Total (%)**	
Not ill Appearing	113 (96)	(92)	3 (3)	(33)	2 (2)	(7)	118 (101*)	(74)
Not Sure	8 (57)	(7)	4 (29)	(44)	2 (14)	(7)	14 (100)	(9)
Ill Appearing	2 (7)	(11)	2 (7)	(22)	23 (85)	(85)	27 (98*)	(17)
Total	123 (77)	(100)	9 (6)	(100)	26 (16)	(100)	159 (99*)	(100)

Table 4A. Intra-rater agreement for first rater

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Table 4B

Second rater Gestalt Impression	Second rater Impression after examining							
	Not ill Appearing (%)		Not Sure (%)		Ill Appearing (%)		Total (%)	
Not ill Appearing	113 (94)	(90)	3 (3)	(38)	4 (3)	(15)	120 (100)	(75)
Not Sure	9 (56)	(7)	5 (31)	(63)	2 (13)	(8)	16 (100)	(10)
Ill Appearing	3 (13)	(2)	0 (0)	(0)	20 (87)	(77)	23 (100)	(14)
Total	125 (79)	(100)	8 (5)	(100)	26 (16)	(100)	159 (100)	(100)

Table 4B. Intra-rater agreement for second rater.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Table 5(on next page)

Table 5. Inter rater reliability

Inter rater reliability measured by Cohen's κ, weighted κ using two commonly employed weighting schemes, polychoric correlation, and Gwet's AC1. * Weights 1-|i-j|/(k-1), ** Weights 1 – [(i-j)/(k-1)]^2 where *i* and *j* index the rows and columns of the ratings by the raters and *k* is the maximum number of possible ratings, (l) linear weighted, (q) quadratic weighted. This table is expanded to include other measures of inter-rater agreement in the appendices.

	Cohen's κ	Weighted κ (*)	Weighted κ (**)	Scott's π	Scott's π(l)	Scott's π(q)	Polychoric correlation	Gwet's AC1	Gwet's AC2(l)	Gwet's AC2(q)
Inter-rater Gestalt	0.119	0.181	0.223	0.118	0.177	0.261	0.334	0.550	0.601	0.635
Inter-rater After exam	0.235	0.283	0.314	0.216	0.261	0.289	0.482	0.655	0.672	0.683
Intra-rater First rater	0.690	0.781	0.844	0.695	0.777	0.833	0.955	0.852	0.893	0.920
Intra-rater Second rater	0.651	0.714	0.758	0.671	0.734	0.777	0.912	0.837	0.871	0.893

Figure 1

Patinet flow through the study

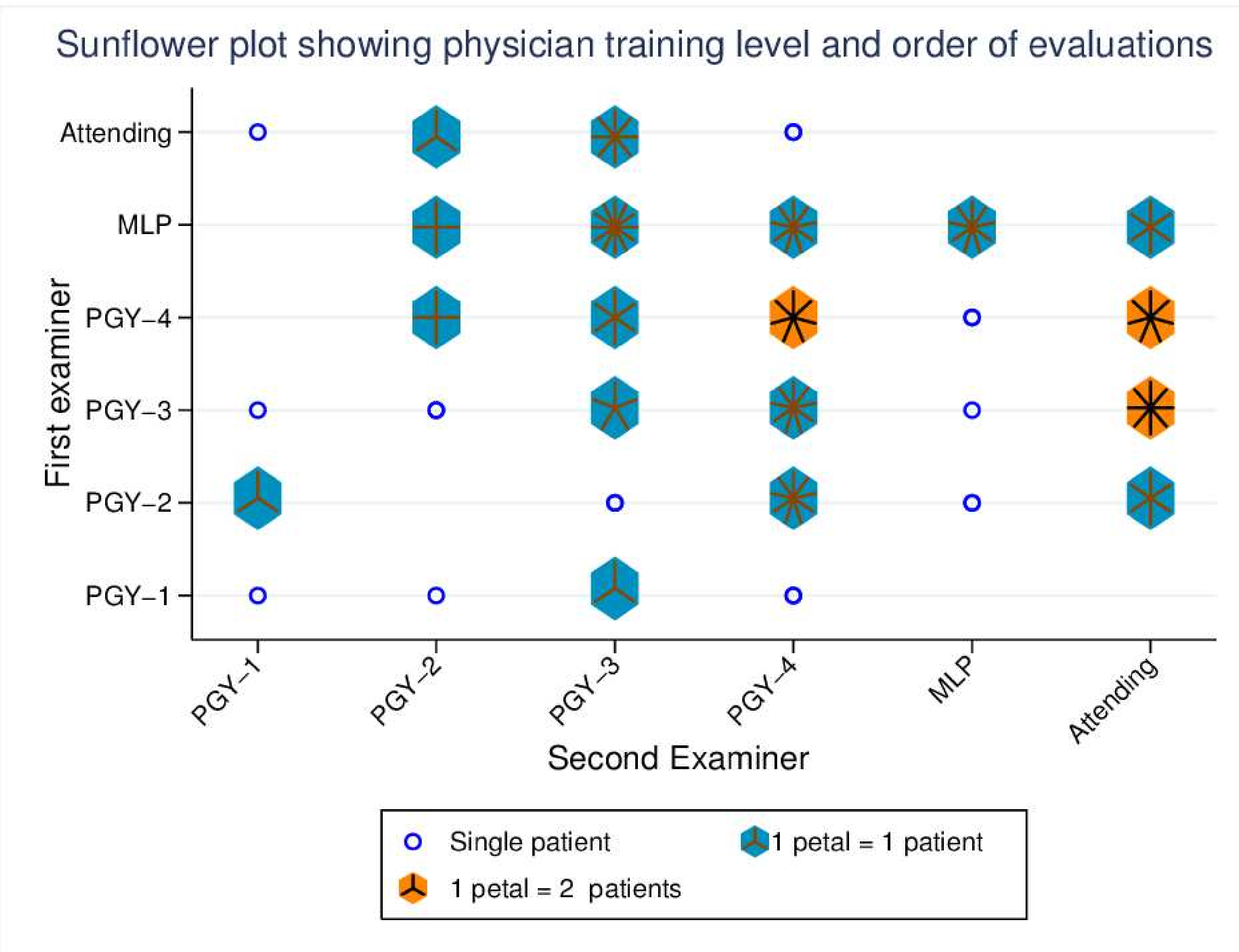
Figure 1. Patinet flow through the study.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Figure 2

Sunflower plot showing trianing level of each provider

Sunflower plot showing physician or provider training level and the order of evaluations. PGY, post graduate year, MLP, mid-level provider. Sunflower plots address the problem of overlapping points on a graph by using ‘flowers’ rather than points. Each flower consists of a number of dark or light petals. Each flower petal represents a number of points. Each blue petal represents one infant; each orange petal represents two.

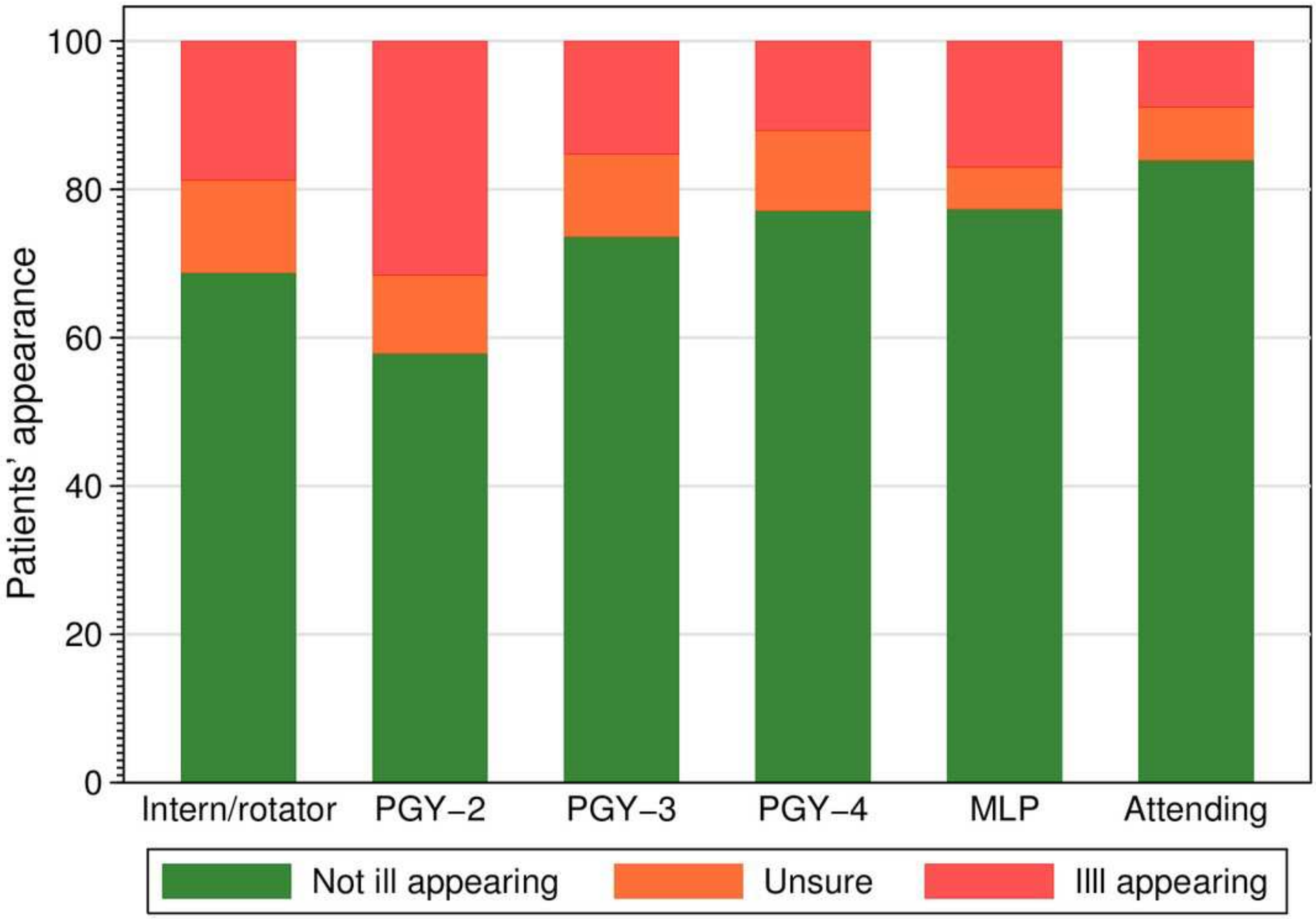


PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Figure 3

Classsification selected and provider training

Frequency of classification selected by provider experience. PGY, post graduate year, MLP, mid-level provider.

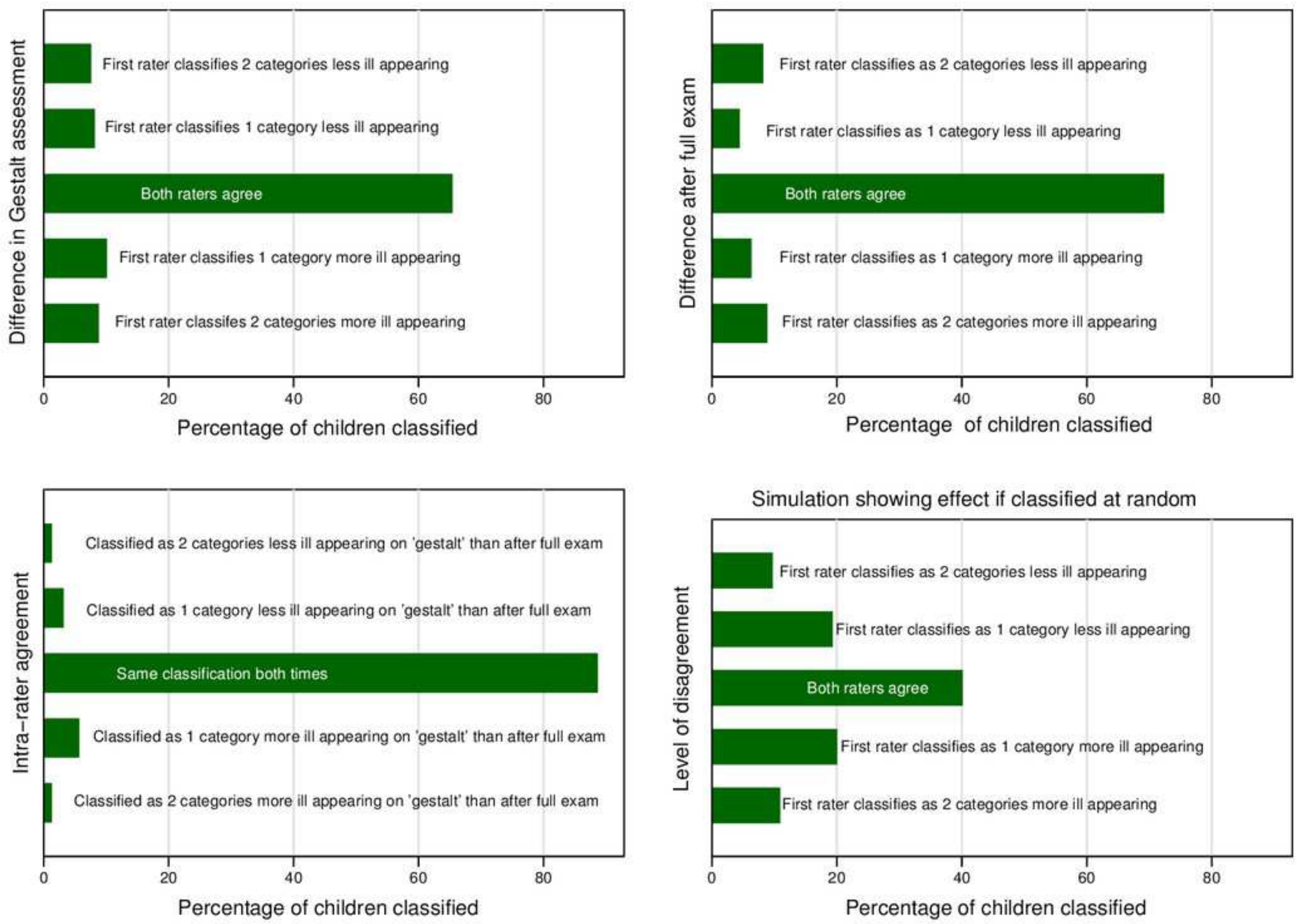


PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul

Figure 4

Graphical analysis of agreement between examiners.

Agreement between examiners' initial 'gestalt' impression, agreement between examiners' after completing their exam, and a simulation showing a uniform random agreement.



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.444v1 | CC-BY 4.0 Open Access | received: 20 Jul 2014, published: 20 Jul