Exhaustive mathematical analysis of simple clinical measurements for childhood pneumonia diagnosis

Keegan Kosasih, Udantha Abeyratne
Brisbane, Australia

Original article

Background: **Pneumonia is the leading cause of mortality for children below 5 years of age. The majority of these occur in poor countries with limited access to diagnosis. The World Health Organization (WHO) criterion for pneumonia is the de facto method for diagnosis. It is designed targeting a high sensitivity and uses easy to measure parameters. The WHO criterion has poor specificity.**

Methods: **We propose a method using common measurements (including the WHO parameters) to diagnose pneumonia at high sensitivity and specificity. Seventeen clinical features obtained from 134 subjects were used to create a series of logistic regression models. We started with one feature at a time, and continued building models with increasing number of features until we exhausted all possible combinations. We used a *k*-fold cross validation method to measure the performance of the models.**

Results: **The sensitivity of our method was comparable to that of the WHO criterion but the specificity was 84%-655% higher. In the 2-11 month age group, the WHO criteria had a sensitivity and specificity of 92.0%±11.6% and 38.1%±18.5%, respectively. Our best model (using the existence of a runny nose, the number of days with runny nose, breathing rate and temperature) performed at a sensitivity of 91.3%±13.0% and specificity of 70.2%±22.80%. In the 12-60 month age group, the WHO algorithm gave a sensitivity of 95.7%±7.6% at a specificity of 9.8%±13.1%, while our corresponding sensitivity and specificity were 94.0%±12.1% and 74.0%±23.3%, respectively (using fever, number of days with cough, heart rate and chest in-drawing).**

Conclusions: **The WHO algorithm can be improved through mathematical analysis of clinical observations and measurements routinely made in the field. The method is simple and easy to implement on a mobile phone. Our method allows the freedom to pick the best model in any arbitrary field scenario (e.g., when an oximeter is not available).**

World J Pediatr 2017;13(5):446-456

Key words: **developing countries;**
diagnosis;
logistic regression modelling;
pneumonia

Introduction

Pneumonia is one of the leading causes of mortality in children under five worldwide. It is estimated that 905 059 children below the age of five died from pneumonia globally in 2013.[1] It accounted for 14% of a total of 6.3 million child deaths around the world that year.[1,2] The United Nations is aware of the issue and, through the Millenium Development Goal (MDG) 4 program, worked with countries globally to reduce the under-five mortality rate by two thirds between 1990 and 2015.[3-5]

Between 1990 and 2013, child pneumonia mortality fell 58% worldwide from 2.2 million to <1 million, and overall global child mortality fell from 12.7 million to 6.3 million, but these reductions were unevenly distributed and outside the worst affected regions.[1,3] An increasing proportion of pneumonia deaths, up to 80% of under-five children, comes from sub-Saharan Africa and South Asia.[5] In 2010, there were 120 million episodes of pneumonia globally, where 14 million cases progressed to severe episodes and 1.3 million cases led to death.[6] Child pneumonia deaths have decreased at a slower pace than other leading causes of child mortality such as measles and diarrhea.[1] This affects the progress of MDG 4 which, based on current trends, will remain as high as 4.4 million cases of under-five mortality in 2030.[2]

Childhood pneumonia also remains a significant burden in the developed world. Up to 2.6 million cases

Author Affiliations: School of ITEE, The University of Queensland, Sir Fred Schonell Drive, St. Lucia, Brisbane, QLD, Australia (Kosasih K, Abeyratne U)

Corresponding Author: Udantha Abeyratne, School of ITEE, The University of Queensland, Sir Fred Schonell Drive, St. Lucia, Brisbane, QLD 4072, Australia (Tel: +61733469063; Fax: +61733654999; Email: udantha@itee.uq.edu.au)

doi: 10.1007/s12519-017-0019-4
Online First March 2017

were estimated across North America, Europe, Australia, New Zealand and Japan each year, of which 1.5 million cases required hospitalization and approximately 3000 children under five years of age succumbed to the disease.[7]

One of the key developments still missing in the global fight against pneumonia is the absence of a rapid, low cost diagnostic method/system.[1,8-14] Diagnosing each case accurately and precisely is difficult even with state of the art equipment, and even more so in poor resource settings. Development assistance targeting pneumonia is relatively scarce and accounted for only 2% of overall global health financing.[1] In 1990, the World Health Organization (WHO) and United Nations International Children's Emergency Fund proposed the WHO criteria for childhood pneumonia classification in resource-poor regions. This is the current de facto diagnostic method used by community health workers in resource limited settings as a rapid low cost alternative in frontline health facilities. Supplementary Table 1 shows the WHO/IMCI (Integrated Management of Childhood Illness) guideline for pneumonia classification in resource poor regions. The WHO/IMCI guideline dictates that if a patient exhibits symptoms of cough/breathing difficulty, the patient is screened for the next step. Breathing rate is taken and if it exceeds the limit [50 breaths per minute (bpm) for age 2-11 months, 40 bpm for age 12-60 months], non-severe pneumonia is declared. Danger signs such as lower chest indrawing and inability to feed or drink would put the patient in the severe pneumonia category requiring immediate attention.

Researchers have generally recognized the limitations of the WHO criteria, which are sensitive but not very specific.[9,15,16] Over the years, others have suggested the addition of fever,[17] grunting and nasal flaring,[8] temperature and oxygen saturation.[18] Rambaud-Althaus et al[8] proposed a combination of signs in a decision tree format to improve clinical diagnosis accuracy.Pneumonia Etiology Research for Child Health investigators developed their own standard interpretations of the symptoms and signs based on the WHO criteria for a clinical case definition of pneumonia.[16]

All these approaches make important contributions to dealing with the global burden of pneumonia, but largely suffer from the same type of limitations afflicting the WHO criteria for resource-poor regions. These methods also rely on health workers to perform measurements and interpret data using basic binary decisions around fixed thresholds.

One method that has yet to be investigated by any group so far, is the use of logistic regression modelling to predict the risk of a child suffering from pneumonia based on a number of simple clinical observations. Logistic regression is one of the most popular models used in medical studies for diagnostic and prognostic purposes. No less than 28 500 studies using logistic regression were found in a 2002 review of publications indexed in Medline.[19] Logistic regression is an efficient and powerful way for predicting binary outcomes based on measuring unique contribution of multiple independent variables.[20] Several studies in the past have used logistic regression to predict risk factors in adult pneumonia patients with positive results.[21,22] One study included children under five in their logistic regression modelling of pneumonia patients, but the aim was directed at analyzing risk factors for mortality instead of initial diagnosis.[23]

One method that has yet to be investigated is the use of mathematical modelling such as logistics regression to diagnose pneumonia based on simple clinical observations. In our previous work, we have demonstrated the benefits of this approach in automatic identification of wet and dry coughs in children,[24] as well as diagnosing childhood pneumonia based on cough sounds.[25]

Our aim is to develop and evaluate models that can increase the specificity of the WHO algorithm while retaining its high sensitivity. In Section II, we describe our patient database as well as the detailed methodology used in our analysis. Section III presents the results of our analysis in conjunction with a discussion of the model performances compared with the WHO criteria. Concluding remarks, future works as well as study limitations can be found in Section IV.

Methods

Study organization

The clinical data used for this study were collected by the Gadjah Mada University-Sardjito Hospital, Yogyakarta, Indonesia, in partnership with The University of Queensland, Brisbane, Australia. The data collection began in December 2010 and continued until March 2014. The ethics committees of the Sardjito Hospital and The University of Queensland approved the study protocol. The inclusion/exclusion criteria are given in Supplementary Table 2. Patients are included if they exhibit any 2 symptoms of cough, sputum, increased breathlessness and temperature >37.5°C. Parental consent was sought prior to inclusion if the patient met the criteria and excluded if consent was not granted. Exclusion criteria also applied to patients showing symptoms of advanced disease, terminal lung cancer and/or requiring a nasal drip IV, as these may skew the outcomes. As a precaution, patients showing droplet-spread disease were also excluded.

Diagnostic definitions

The reference diagnosis used in this study is the overall

diagnosis provided by the pediatricians on the basis of clinical presentation, laboratory tests, chest X-ray, and the clinical course of the disease. An X-ray was performed only on subjects clinically suspected of pneumonia and, on other occasions, where there is clear need for it. Therefore, not all our subjects underwent a chest X-ray.

Study protocol

All children who satisfied the inclusion criteria were invited to participate in the study. Each child's history and clinical measurements were recorded as part of the hospital admission process. Diagnostic outcomes and all test results collected from the subjects in the course of normal diagnosis/management of the disease were made available to this study. Supplementary Table 3 lists some of the information recorded which was used for analysis in this paper. The test parameters included the existence of fever, cough, breathing difficulty, runny nose, and chest indrawing as a binary yes/no observation. It also included the following data as numbers: age, weight, height, breathing rate, temperature, body mass index, oxygen saturation, and number of days suffering fever, cough, breathing difficulty, runny nose. Other diagnostic measures such as blood/sputum analysis and chest X-ray were performed only if the attending physician deemed it to be necessary.

Study population

We recruited 222 children in total: 93 females, 129 males with a median age of 9 months and an inter-quartile range of 4.25-20 months. Our population came from subjects admitted to the hospital ward. Our intention was to focus on the clinical parameters of interest in diagnosing pneumonia in resource poor regions. The dataset comprised 134 children with the complete list of parameters specified earlier. We excluded 88 patients from further consideration due to the absence of one or more of the required parameters. The distribution of the chosen 134 children closely represented the initial 222 children recruited, as shown in Supplementary Table 4. There were 71 and 63 children in the 2-11 months and 12-60 months age groups, respectively. Of the 134 children, 96 were diagnosed with pneumonia, whereas the remaining 38 were a mix of asthma, bronchitis, bronchiolitis, heart disease, malnutrition, wheezing, etc. The non-pneumonia group served as the control set for this study.

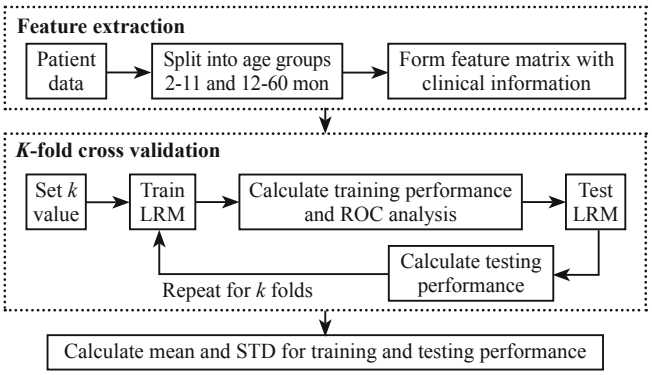

Fig. 1. Flow diagram for the machine learning process of training a logistic regression model (LRM) and testing the performance of each selected features in diagnosing pneumonia. ROC: receiver operating characteristic; STD: standard deviation.

Analysis of data

The flow diagram presented in Fig. 1 details the process used in analyzing our data. The data set is split into two age groups, 2-11 months and 12-60 months. Clinical features from each group are tabled into a feature matrix for processing.

Using a *k*-fold cross validation method, each age group was randomly split into k number of folds. An iterative process was then adopted in which one fold of data was retained as the testing set whilst the rest of the data was used for training a logistic regression model (LRM). A good general explanation of the logistic regression method used in medical applications can be found in a paper by Sainani.[26]

The LRM outputs the probability of the existence of pneumonia based on the specified predictors, to which a cut-off threshold is applied to make the output a binary decision. This threshold was carefully selected following a receiver operating characteristic (ROC) analysis to separate the positive and negative pneumonia cases as cleanly as possible.

In the LRM design, we commenced by using one feature at a time and computing the performance of the resulting models. We then exhaustively searched all combinations of two features taken at a time. This process was continued until we reached all 17 features taken at a time. In each iteration, the trained models were evaluated according to their sensitivity (S_n), specificity (S_P), accuracy (A_{cc}), and the area under the curve (AUC). AUC was only available for the training data to set the diagnostics threshold.

Note that in each fold of the *k*-fold cross validation, the data set was divided into non-overlapping training and testing sets, and the performance was estimated separately for both the training and testing sets. Each iteration generated *k* number of ROC curves and *k* sets of training and testing performance measures. Each iteration also generated *k* number of trained LRM models. The trained LRM models were used to calculate the performance of the training set. The LRM models were then fixed, and used on the testing data set to compute the testing performance and validate the trained model. Each set of LRM models was considered

final for its respective fold. Hence, for testing performance, no AUC data were available.

The best performing models were chosen based on the means and standard deviations (SDs) of the training and testing performances. These numbers were calculated and are reported in the Results section. Similarly, the WHO criteria performance numbers were calculated for the testing sets and represented using their means and SDs. We then used the performance values to determine which parameters, in which combinations could provide the best diagnostic outcomes with the testing sets.

This process was iterated for each feature combination used. First we analyzed the LRM performance of using one, two, and three features at a time. Next, we exhaustively analyzed all possible feature combinations with up to 17 features being used at once. Supplementary Table 5 shows the possible combinations for each number of features used in the creation of the LRM model. Using one feature at a time gives 17 possible combinations, whereas using all 17 features at a time would have only one possible combination. The number of possible combinations rises significantly in between. For example, the use of 8 features at a time results in 24 310 combinations. In total, the number of models tested in this study is comprised of 131 071 combinations. Given the large number of models tested, the ROC curve analysis to find the best cut-off threshold for each model becomes very important. We selected the threshold targeting a $S_n \geq 90\%$ with S_p as high as possible. This also had the benefit of lowering the false discovery rate. As we mentioned earlier, our aim is to improve the S_p of the WHO algorithm, while maintaining high S_n.

Results

In this section, we show the results of our analysis, starting from the cross-validation process and the WHO/IMCI algorithm performance in our patient groups. We then describe the performance of our models and compare with the WHO outcomes in the 2-11 month age group, followed by the 12-60 month age group.

The cross-validation technique

As detailed in Methods section, we used k-fold cross-validation to train and evaluate our classifier models. In this study we set k=8, resulting in 8-9 children in each fold. Higher k values, such as the more commonly used k=10, would result in 6-7 children in each fold. We deemed this number as insufficient and decided k=8 gives better balance for the testing data. Note that in each fold of the cross validation, training and testing sets are mutually exclusive, that is training and test testing sets do not overlap.

WHO/IMCI performance

We applied the WHO criteria (Supplementary Table 1) to data in each fold of the k-fold cross validation data set, and computed the mean and the SD across all folds. Results are shown in Table 1. As expected, WHO criteria yielded high S_n with relatively small SD across both age groups, but at a poor specificity S_p.

Our target is to maintain the high sensitivity of the WHO algorithm while increasing the specificity. Next we describe the performance of the proposed method. As the analysis was done separately for each age group, we will begin by presenting the results for the 2-11 month age group.

Performance in the 2-11 month age group

Supplementary Fig. 1 shows mean S_n and S_p values for models using one, two, and three features at a time for the 2-11 month age group. Our feature set consisted of 17 observations/measurements as listed in Supplementary Table 3. The number of total combinations of one-feature taken at a time is 17, leading to 17 LRM models with one feature as the input (Supplementary Fig. 1, top frame). Similarly, two features at a time and three features at a time give us 136 and 680 LRM models respectively (Supplementary Fig. 1, middle and bottom frames). Overall, significant benefits were found in combining features up to four at a time in one LRM.

Fig. 1 shows the ROC curve analysis for 12 trained LRM models we selected for further consideration. Two models each are from single and double feature combinations, five from triple feature combinations, and three with four feature combinations. The solid line in each frame represents the mean ROC curve, formed over the k folds. Crosses on each frame represent the SD of the S_n and 1-S_p at points shown. On each frame of Fig. 2, we also graphically illustrate (see boxes) the performance possible with the WHO/IMCI algorithm for resource-limited regions. The center of the box indicates the mean performance, and the height and width of the box represent SD. Table 2 shows the training and testing performance numbers for the 12 models.

When a single feature is used to create the LRM for the 2-11 month age group, the best features in terms of testing performance were breathing rate (S_n of 91% and S_p of 35%) and chest in-drawing (S_n of 98% and S_p of 33%). These numbers closely matched the performance

Table 1. WHO criteria as applied to k-fold testing sets

Age group	Classification performance (%), mean±standard deviation				
	Sensitivity	Specificity	Accuracy	PPV	NPV
2-11 mon	92.0±11.6	38.1±18.5	69.1±11.2	67.6±12.5	76.2±28.0
12-60 mon	95.7±7.6	9.8±13.1	66.5±12.8	66.5±15.2	60.0±49.0

PPV: positive predictive value; NPV: negative predictive value.

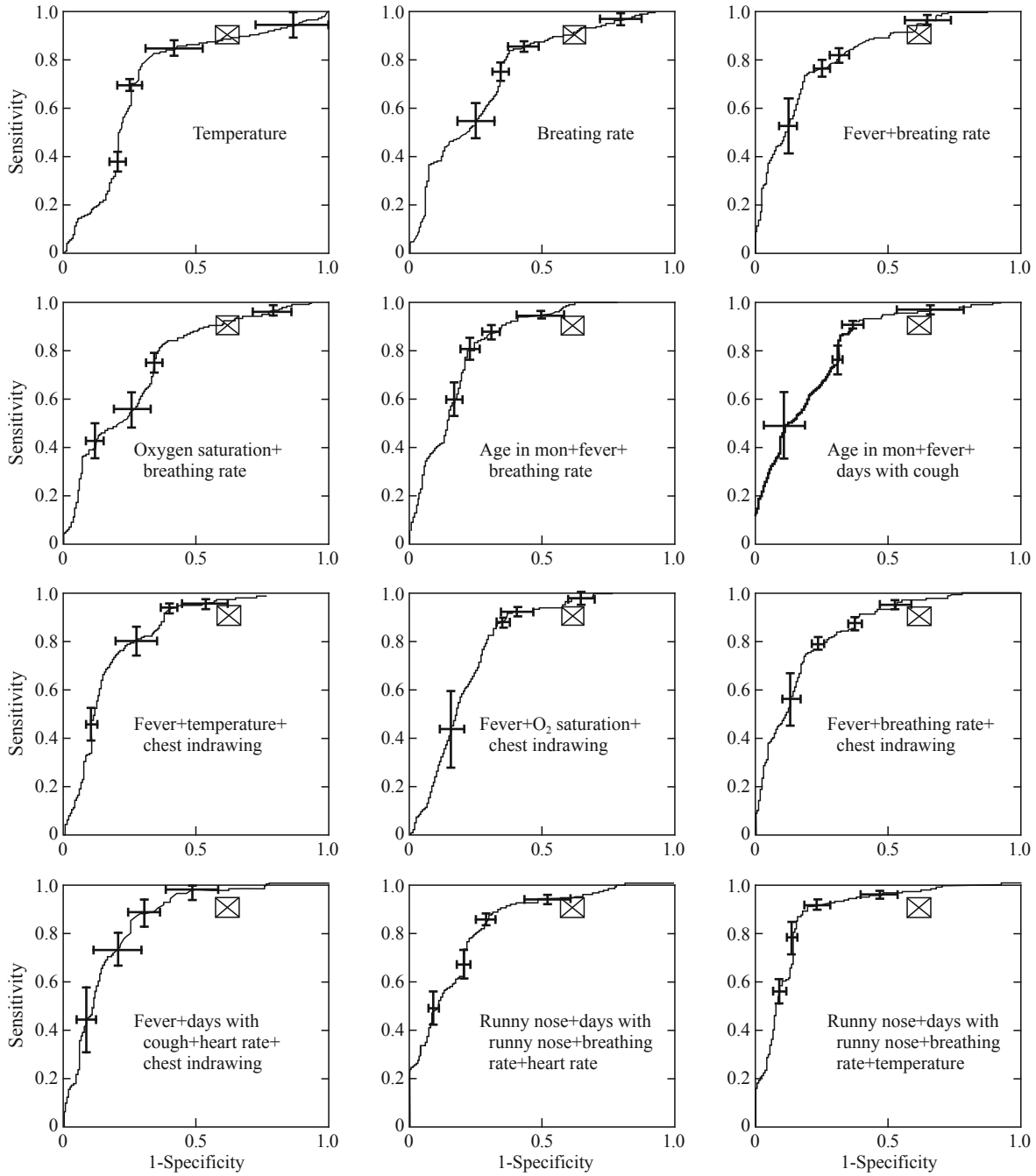

Fig. 2. ROC curve analysis for the age group of 2-11 month. The solid line in each frame represents the mean ROC curve, formed over *k*-iterations. Crosses on each frame represent the SD of the *Sn* and 1-*Sp* at points shown. On each frame, the performance of WHO/IMCI algorithm for resource-limited regions is graphically illustrated (see boxes). The center of the box indicates the mean performance, and the height and width of the box represent SD. ROC: receiver operating characteristic; SD: standard deviation; *Sn*: sensitivity; *Sp*: specificity; WHO: World Health Organization; IMCI: integrated management of childhood illness.

of the WHO criteria, as it also relies on the same features for childhood pneumonia classification.

Individually, the breathing rate and temperature exhibit the highest AUC in the training performance. However, the temperature model shows higher S_n and lower S_p compared with the WHO criteria, as opposed to breathing rate model which has comparable numbers. On the testing dataset, both models demonstrate high S_n with little SD, but the SDs of the S_p vary wildly, rendering both models unusable by themselves. This suggests that the WHO criteria are still more reliable when compared to single feature LRM models.

The use of two features at a time boosts the S_p to 50% for certain feature combinations while maintaining

S_n around 90%. This is a significant improvement from the S_p of the best single feature model. The best performers are models using breathing rate with oxygen saturation, and, breathing rate with fever. Both feature combinations exhibit high AUC (75%-82%) in training.

We continued to add features further until the optimal LRM feature combinations were found. On the three simultaneous feature models, the overall testing performances are higher than the double feature ones. Mean S_n levels remain largely the same around 90% and mean S_p levels are on average 30% higher than the double feature models. The SD level for S_n is unchanged but for S_p is 44% smaller. The best performing model for this category includes fever, oxygen saturation and chest in-drawing as parameters, achieving an S_n value of 88.1%±13.6% and an S_p value of 61.9%±8.7%. Compared with the WHO/IMCI algorithm (S_n and S_p of 92.0%±11.6% and 38.1%±18.5%, respectively), the mean S_n is 4% lower while S_p is 62% higher. The SDs are 17% larger for S_n and 53% smaller for S_p compared with WHO. Thus, the best triple feature model performs much better than WHO criteria in terms of S_p, with a small loss of S_n.

Further improvements in classification performance are found using four features at a time. The best performing model uses the existence of runny nose, number of days with runny nose, breathing rate and temperature (91.3%±13.0% S_n and 70.2%±22.80% Sp). The mean S_n is on a par with the WHO results, and the mean S_p is 84% higher. The SD for both S_n and S_p are, however, slightly higher compared with the WHO/IMCI algorithm. The second best performing model uses runny nose, days with runny nose, breathing rate, and heart rate at S_n of 91.5%±9.2% and S_p of 66.0%±26.3%. The mean S_n is also on a par with the WHO results while S_p is 73% higher. For SDs, they are 20% smaller for S_n and 42% larger for S_p compared with WHO.

Table 2. Performance comparison between various models for diagnosing pneumonia in children aged 2-11 months

Features	Training performance (%)					Testing performance (%)		
	S_n	S_p	A_{cc}	AUC	CO	S_n	S_p	A_{cc}
Temperature								
Mean	93.2	11.9	59.9	71.7	38.4	88.8	11.3	58.7
SD	1.4	9.8	3.9	2.5	3.7	13.0	12.2	12.2
Breathing rate (BR)								
Mean	93.5	35.4	69.8	74.9	35.6	91.3	35.2	65.7
SD	1.5	12.0	4.3	2.2	2.5	13.0	29.8	17.1
Chest indrawing								
Mean	97.6	34.4	71.8	66.0	67.9	97.9	32.9	71.6
SD	1.0	3.4	1.6	1.8	1.8	5.9	23.7	10.7
Fever+breathing rate								
Mean	93.2	44.5	73.2	82.1	33.5	86.0	50.2	69.1
SD	1.5	11.2	4.7	2.6	3.7	17.9	26.6	18.2
Oxygen saturation+breathing rate								
Mean	91.8	37.9	69.8	75.0	37.2	91.2	35.2	65.7
SD	0.3	11.6	5.0	2.3	3.4	13.0	29.8	17.0
Age (mon)+fever+breathing rate								
Mean	91.8	63.7	80.3	83.3	39.9	86.0	66.7	76.6
SD	0.3	7.9	3.4	2.4	5.1	17.9	22.5	16.8
Age (mon)+fever+days with cough								
Mean	91.8	63.7	80.3	81.2	41.3	83.5	67.3	76.1
SD	0.3	4.5	2.4	1.9	10.6	14.5	23.9	18.5
Fever+temperature+chest indrawing								
Mean	91.8	62.1	79.7	82.6	46.0	90.6	57.7	77.8
SD	0.3	1.2	0.8	1.8	5.6	13.7	13.0	10.6
Fever+oxygen saturation+chest indrawing								
Mean	92.8	64.0	81.1	77.9	44.1	88.1	61.9	77.6
SD	1.5	3.0	1.6	1.9	8.7	13.6	8.7	10.9
Fever+breathing rate+chest indrawing								
Mean	91.8	59.2	78.5	83.6	41.4	86.0	56.3	73.6
SD	0.3	4.7	2.2	2.1	4.3	17.9	10.1	14.1
Fever+days with cough+heart rate+chest indrawing								
Mean	92.2	65.9	81.5	84.7	59.6	86.0	63.5	76.3
SD	1.0	6.5	2.6	2.6	7.9	15.5	18.6	13.3
Runny nose+days with runny nose+BR+temperature								
Mean	91.8	77.4	85.9	88.1	51.0	91.3	70.2	82.0
SD	0.3	6.1	2.5	2.1	6.1	13.0	22.8	9.3
Runny nose+days with runny nose+BR+heart rate								
Mean	91.8	64.0	80.5	83.5	41.5	91.5	66.0	80.1
SD	0.3	4.6	2.0	2.2	2.6	9.2	26.3	12.1

AUC: area under the curve; CO: cut-off threshold; SD: standard deviation; S_n: sensitivity; S_p: specificity; A_{cc}: accuracy.

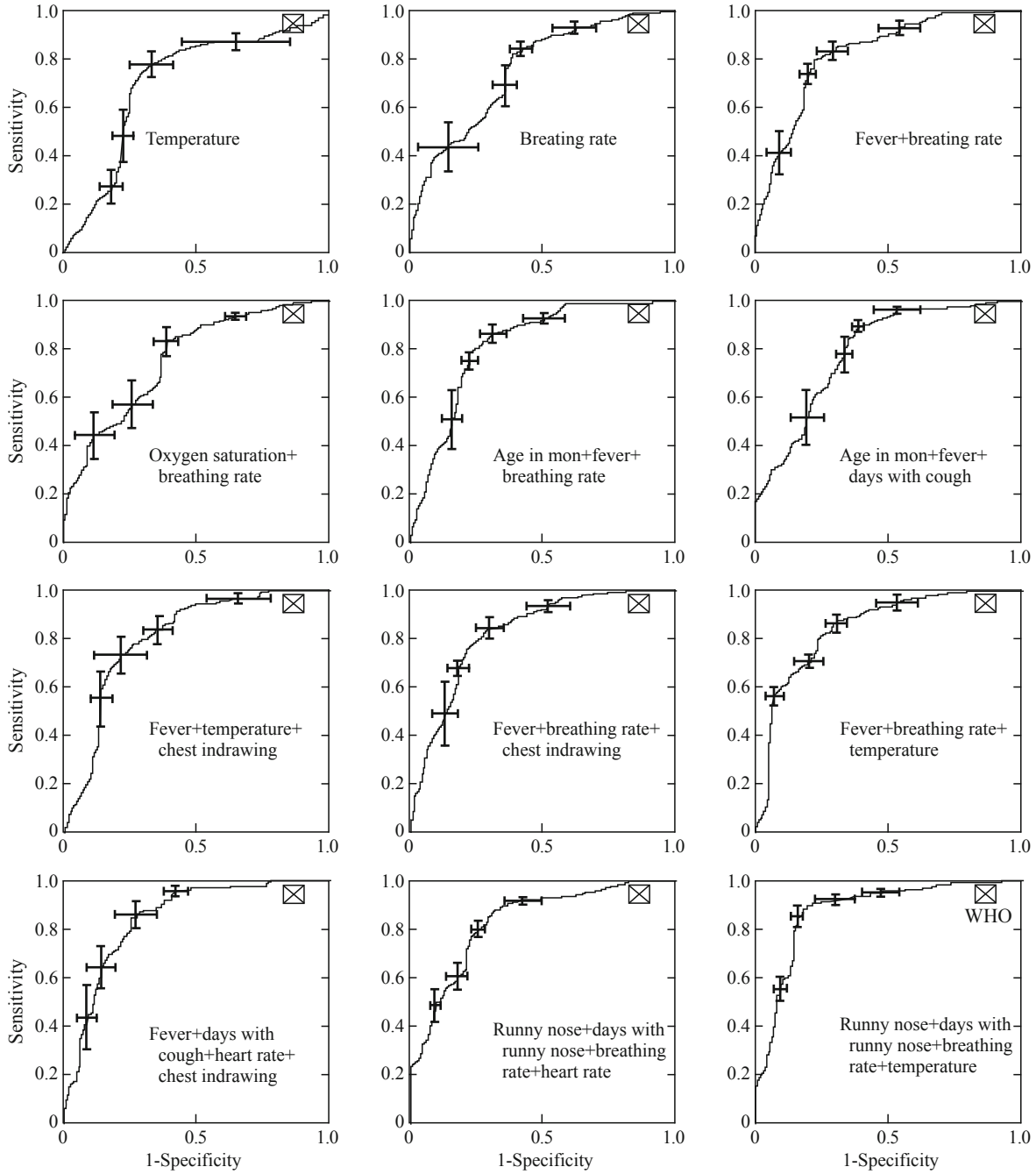

Fig. 3. ROC curve analysis for the age group of 12-60 month. The solid line in each frame represents the mean ROC curve, formed over *k*-iterations. Crosses on each frame represent the SD of the S_n and 1-S_p at points shown. On each frame, the performance of WHO/IMCI algorithm for resource-limited regions is graphically illustrated (see boxes). The center of the box indicates the mean performance, and the height and width of the box represent SD. ROC: receiver operating characteristic; SD: standard deviation; S_n: sensitivity; S_p: specificity; WHO: World Health Organization; IMCI: integrated management of childhood illness.

Performance in the 12-60 month age group

For the 12-60 month age group, the same process is repeated, starting with observation of the ROC curves from 12 trained LRM models chosen for comparison, as shown in Fig. 3. Table 3 shows the training and testing results for LRM models in the 12-60 month age group. The two best performing models for both age groups are compared in Supplementary Fig. 2.

For the single feature category, breathing rate and chest in-drawing (individually) still exhibit the best performance in general. The WHO/IMCI algorithm implementation for this age group demonstrates S_n of 95.7%±7.6% and S_p of 9.8%±13.1%. The best double feature LRM models in this age group also include breathing rate as a parameter.

The best two models are breathing rate with fever and breathing rate with oxygen saturation.

For triple features, the best testing performance was observed when using fever, temperature, and chest in-drawing. This combination reached a S_n of 92.1%±15.8% and S_p of 51.3%±39.0% for testing performance. The mean S_n is still comparable to the WHO criteria with 3% disparity, but the mean S_p of the LRM model is 423% higher. The SD of the S_n is 107% larger compared with WHO, and for S_p it is 197% greater. The best S_p in this category is found in the combination of fever, breathing rate and temperature with S_p of 74.8%±30.2%. This is a 663% increase of mean S_p over WHO with 130% increase in SD. The mean S_n value is also 8% lower compared to WHO results while SD remains 100% higher.

In models with four features, the best performing model uses the existence of fever, number of days with cough, heart rate, and existence of chest in-drawing (S_n of 94.0%±12.1% and S_p of 74.0%±23.3%). The mean S_n is 2% lower than the WHO performance and the mean S_p is 655% higher. The second best performing model utilizes runny nose, days with runny nose, breathing rate, and temperature with S_n and S_p of 91.4%±12.1% and 71.9%±36.4%, respectively.

Recurrent features in best performing models

Following the recurrent appearance of certain features amongst the best performing LRMs in all feature combinations, we decided to systematically explore these features in order to rank the most significant features out of the 17 considered.

We set a threshold S_n of 90% and S_p of 70% on the mean testing performance for all possible combinations, from using one feature at a time to 17 features, and found 20 models that meet the criteria (eight models from the 2-11 month age group and 12 from the 12-60 age month group, respectively). Table 4 shows the number of recurrence for each feature within the top 20 feature combinations. Several measurements such as the breathing rate and observations such as the existence of runny nose are dominantly present as recurrent features in good models.

Table 3. Performance comparison between trained models and WHO criteria for children aged 12-60 months

Features	Training performance (%)					Testing performance (%)		
	S_n	S_p	A_{cc}	AUC	CO	S_n	S_p	A_{cc}
Temperature								
Mean	93.0	10.7	64.4	70.4	46.6	88.0	19.2	61.6
SD	1.6	13.6	3.9	2.6	3.2	14.0	35.0	35.0
Breathing rate (BR)								
Mean	93.0	41.5	75.1	74.9	42.6	88.2	45.2	71.2
SD	1.6	9.4	2.9	2.8	4.2	14.6	32.0	15.2
Chest indrawing								
Mean	97.6	32.0	74.6	64.8	72.3	97.5	32.3	74.3
SD	1.0	4.9	3.0	2.5	3.0	7.1	35.8	20.6
Fever+breathing rate								
Mean	92.0	49.6	77.3	82.0	41.4	88.3	44.2	67.9
SD	1.1	10.5	3.1	2.4	5.9	15.2	47.7	21.1
Oxygen saturation+breathing rate								
Mean	91.6	47.0	76.0	79.3	46.5	88.2	57.7	72.2
SD	0.4	6.5	3.2	2.7	6.5	14.6	31.3	17.5
Age (mon)+fever+breathing rate								
Mean	91.6	58.4	80.0	81.5	46.9	88.3	70.6	75.9
SD	0.4	10.3	3.6	2.5	7.2	15.2	33.5	14.9
Age (mon)+fever+days with cough								
Mean	91.6	60.4	80.7	79.3	53.5	84.3	59.6	74.6
SD	0.4	4.3	2.0	2.2	9.4	21.2	37.6	13.4
Fever+temperature+chest indrawing								
Mean	92.0	57.7	80.0	80.6	51.8	92.1	51.3	79.2
SD	0.9	3.8	1.4	2.6	9.7	15.8	39.0	16.5
Fever+breathing rate+chest indrawing								
Mean	91.6	56.4	79.4	82.3	49.0	88.3	58.1	74.3
SD	0.4	9.6	3.3	2.4	9.1	15.2	39.2	17.0
Fever+breathing rate+temperature								
Mean	92.0	61.1	81.2	85.5	42.1	87.9	74.8	77.5
SD	0.9	7.9	3.0	1.7	5.1	15.4	30.2	16.9
Fever+days with cough+heart rate+chest indrawing								
Mean	91.6	68.3	83.4	82.7	62.1	94.0	74.0	84.4
SD	0.3	6.6	2.7	3.4	7.7	12.1	23.3	8.8
Runny nose+days with runny nose+BR+temperature								
Mean	91.6	74.0	85.5	87.5	52.3	91.4	71.9	87.5
SD	0.3	4.9	1.7	2.8	9.8	12.1	36.4	9.2
Runny nose+days with runny nose+BR+heart rate								
Mean	91.6	61.8	81.2	84.0	47.6	85.9	59.4	74.6
SD	0.3	3.2	1.6	2.8	3.3	21.2	37.6	17.7

AUC: area under the curve; CO: cut-off threshold; SD: standard deviation; S_n: sensitivity; S_p: specificity; A_{cc}: accuracy.

Table 4. Number of feature occurrence in the models showing >90% sensitivity and >70% specificity

No.	Feature name	2 to 11 mon	12 to 60 mon
1	Age in mon	0	0
2	Fever	0	2
3	Days with fever	0	0
4	Cough	0	0
5	Days with cough	0	2
6	Runny nose	8	9
7	Days with runny nose	8	8
8	Breathing difficulty	0	0
9	Days with breathing difficulty	0	0
10	Weight	0	1
11	Height	0	0
12	Breathing rate	8	8
13	Heart rate	0	1
14	Temperature	8	10
15	Body mass index	4	4
16	O_2 saturation	4	4
17	Chest indrawing	4	6

Discussion

One particular aim of this study was to explore if common clinical observations and measurements could be utilized to diagnose pneumonia at specificities higher than possible with the WHO algorithm, while maintaining the sensitivity of at least 90%. Our results have illustrated that this is indeed possible. Our best performing models demonstrated a sensitivity of 91% while achieving an S_p in the range of 70%-72% depending on the age of the subjects. These numbers represented 84%-655% increase in S_p compared to the WHO/IMCI algorithm, which had S_p ranging between 10%-38%. Our results are based on k-fold cross validation, and the reported outcomes are thus not on the same data used to train a particular model.

The number of clinical observations and measurements needed to achieve a desired performance provides useful insight in designing clinical protocols targeting resource-poor areas. Results we obtained indicated that our single feature models perform similar to the WHO/IMCI algorithm. Addition of second, third and fourth features significantly improve S_p while S_n continues to hold above 90%. Beyond four features, the calculation complexity rises without any performance gain.

One important contribution of this paper is the identification of most important clinical features and measurements that may substantially increase the accuracy of diagnosing pneumonia in resource-poor regions. We surveyed our exhaustive model database for the repeated appearance of features in models satisfying S_n>90% and S_p>70%.

The breathing rate appeared as a feature in 16 models across both age groups out of a total of 20. Oxygen saturation and chest indrawing too were important parameters appearing respectively in 8 and 12 models. The significance of these measurements are well known among the medical and research communities. Our work uncovered two parameters of potential significance; "the existence of runny nose" and the "number of days with runny nose", and both of which appeared in 16 out of 20 models, just like the breathing rate. The "existence of fever" also presented as a frequent parameter (4 out of 20 models) for the age group 12-60 months.

Breathing rate is the main measurement used in the WHO/IMCI algorithm to diagnose pneumonia. While it appears an easy parameter to measure, it has been found difficult to achieve in resource-poor regions. Therefore, a major fraction of the global pneumonia diagnosis resources are allotted to improving technologies and protocols to measure the breathing rate.[27-29] Without a reliable breathing rate measurement, the WHO/IMCI methods cannot be used in the field.

Our results suggest that while breathing rate is an important parameter, it is not essential to diagnose pneumonia. For instance, our model using the four features age, existence of fever, existence of cough and days of cough was capable of S_n=83.5%±14.5% and S_p=67.3%±23.9% respectively, for the 2-11 month age group. In the other age group, this model exhibited S_n and S_p of 91.7%±17.8% and 51.0%±34.6%, respectively. Among two-feature models, the combination using fever and days with cough resulted in S_n and S_p of 90.2%±14.0% and 44.3%±25.9%, respectively, for the 2-11 month age group. For the older age group, the model performed with S_n=85.5%±15.2% and S_p=41.9%±47.7%. These results are parallel to our previous observation that breathing rate may not add additional value when mathematical features derived from cough sounds are available for diagnosing pneumonia.[30]

Recently there has been a renewed interest in the use of pulse oximetry in reducing childhood pneumonia mortality in resource-poor settings.[31-33] Hypoxemia is a diagnostic indicator for severe pneumonia and swift access to oxygen treatment could improve the prognosis, when available. In our exhaustive model building process, we found 8 of the 20 best models included oximetry as a feature. Oximetry can be a highly useful feature. However, our results suggest that we can substitute, in its place, simpler feature combinations when a pulse oximeter is not available in the field.

The WHO/IMCI criteria for resource-poor regions have been designed to be highly sensitive to detect pneumonia. Sensitivities such as 94% for those aged <24 months, 62% for ≥24 months have been reported.[17] A high number of false positive results also occur, reducing the specificity of the method (16%-20%).[17] In our previous works on children, we have seen WHO/IMCI performing at a sensitivity of 83% and a specificity of 47% (n=91).[25,30] The WHO/IMCI criteria works well when applied by doctors in conjunction with clinical and radiological analysis, giving performances of 77%-81% sensitivity and 77%-80%

specificity.[34] These numbers are comparable with what we obtained in this paper, though our method did not use laboratory or radiological measurements.

High false positive rate of the WHO criteria can lead to rising antimicrobial resistance in communities and render antibiotics ineffective. It also wastes rare drug stocks and delays early treatment opportunities for diseases with symptom overlap (e.g., malaria).[10,35] In low resource settings where only WHO/IMCI criteria are available, as many as 30% of cases had symptoms compatible with both malaria and pneumonia, necessitating dual treatment.[36] One of these treatments could be redundant. The method presented in this paper could potentially help with these issues by producing more accurate results, even in the absence of key parameters such as breathing rate.

The approach we took in this paper is unique. We systematically exhausted all possible feature combinations in our set of 17 features. Altogether we built and tested 131 071 models, each using different feature combinations. In the literature there are instances where WHO/IMCI algorithm was augmented with one or two other handpicked clinical features (e.g., fever, oximetry) targeting manual interpretation. For instance, Cardoso et al in their 2010 study[17] added fever to WHO/IMCI algorithm and illustrated the specificity increased up to 44% (age group <24 month) and 50% (age group 24-60 months). However, the sensitivity was reduced below that of WHO/IMCI. In particular, in the age group 24-60 months, neither the original WHO/IMCI nor the modified method could achieve sensitivity above 62%. The method we proposed can achieve a sensitivity above 90% while maintaining the specificity at the range 70%-72%. No manual interpretation of features is necessary, and our method can be the basis of a decision device.

After this manuscript was submitted, in an independent development, Naydenova et al[37] published results on a method of combining several features using a machine learning approach. They reported oxygen saturation, temperature, breathing rate and heart rate as leading to the best performance in their model (sensitivity 96.6%, specificity 96.4%). In our work, the same feature combination resulted in an inferior performance (sensitivity 88.8%, specificity 40% in the age group 2-11 months; sensitivity 82.7% and specificity 35.4% in the age group 12-60 months).

One critical difference between our method and the one by Naydenova et al[37] is that they used healthy people as control subjects while we used children with respiratory symptoms satisfying inclusion criteria as our control subjects. Our control subjects were children who visited the hospital seeking treatment for illnesses with symptoms shared with pneumonia, but the medical diagnosis was they had different diseases. The research problem we explored was completely different from the one examined by Naydenova et al[37] and the results are thus not comparable. Separating normal children from pneumonia subjects is a much simpler problem compared with identifying pneumonia subjects from a group of children with a range of respiratory illnesses.

Another critical difference was the reference standard used to diagnose pneumonia. The outcomes of Naydenova algorithms were compared against WHO/IMCI algorithm as the reference standard, which has high sensitivity but poor specificity. Our reference standard used the diagnosis by pediatricians aided with clinical examinations, auscultation and laboratory and radiological results as deemed necessary for a clinical decision. The clinical course of the disease too was considered in the final diagnosis.

In conclusion, we investigated the problem of diagnosing pneumonia in a cohort of pediatric patients visiting a hospital presenting with respiratory complaints. We systematically and exhaustively examined combinations of clinical features in their performance in diagnosing pneumonia. Altogether we built and tested 131 071 LRM, each using different feature combinations. The LRM models we developed could retain the high sensitivity of the WHO/IMCI algorithm while increasing its mean specificity by 84% for the 2-11 month age group and 655% for the 12-60 month age group.

This study was limited by the number of subjects (n=134) used and the reference method used to diagnose pneumonia. The reference standard used in this study is the overall clinical diagnosis aided by auscultation, laboratory analysis and radiography (when deemed clinically necessary by the attending physician) and the clinical course of the subject's response to treatment. Due to the need to limit radiation exposure to children, X-ray imaging was not performed on all subjects in the study.

Funding: This work is partly supported by the Bill & Melinda Gates Foundation, USA, under a Grand Challenges in Global Health Exploration Grant (#OPP1008199 GCE) to Abeyratne. The funding was used for data collection.
Ethical approval: This study was approved by Medical Ethics Research Committees of The University of Queensland, Australia and Gadjah Mada University, Indonesia (approval #s: 2010000338 and KE/FK/7/9/EC, respectively).
Competing interest: None.
Contributors: Kosasih K contributed to the concept and design (35%), interpretation of data (50%), computation (100%), drafting and writing (65%). Abeyratne U contributed to concept and design (65%), interpretation of data (50%), drafting and writing (35%).

References

1 Fullman N, Lim S, Dieleman J, Greenslade L, Graves C, Huynh C, et al. Pushing the pace: progress and challenges in fighting

childhood pneumonia. Seattle, WA: IHME, 2014.
2 Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, et al. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet 2015;385:430-440.
3 We can end poverty: millenium development goals and beyond 2015, 2015. http://www.un.org/millenniumgoals/childhealth.shtml (accessed December 28, 2016).
4 Bryce J, Black R, Victora C. Millennium development goals 4 and 5: progress and challenges. BMC Med 2013;11:225.
5 UN. The millenium development goals report 2014. New York: UN, 2014.
6 Walker CL, Rudan I, Liu L, Nair H, Theodoratou E, Bhutta ZA, et al. Global burden of childhood pneumonia and diarrhoea. Lancet 2013;381:1405-1416.
7 Madhi SA, Wals PD, Grijalva CG, Grimwood K, Grossman R, Ishiwada N, et al. The burden of childhood pneumonia in the developed world: a review of the literature. Pediatr Infect Dis J 2013;32:e119-e127.
8 Rambaud-Althaus C, Althaus F, Genton B, D'Acremont V. Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. Lancet Infect Dis 2015;15:439-450.
9 Qazi S, Were W. Improving diagnosis of childhood pneumonia. Lancet Infect Dis 2015;15:372-373.
10 WHO. Antimicrobial resistance: global report on surveillance 2014. Geneva: World Health Organization, 2014.
11 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.
12 Chang AB, Ooi MH, Perera D, Grimwood K. Improving the diagnosis, management, and outcomes of children with pneumonia: where are the gaps? Front Pediatr 2013;1:29.
13 Esposito S, Principi N. Unsolved problems in the approach to pediatric community-acquired pneumonia. Curr Opin Infect Dis 2012;25:286-291.
14 Rudan I, El Arifeen S, Bhutta ZA, Black RE, Brooks A, Chan KY, et al. Setting research priorities to reduce global mortality from childhood pneumonia by 2015. PLoS Med 2011;8:e1001099.
15 WHO. Consultative meeting to review evidence and research priorities in the management of ARI. Geneva: World Health Organization, 2004.
16 Scott JAG, Wonodi C, Moïsi JC, Deloria-Knoll M, DeLuca AN, Karron RA, et al. The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study. Clin Infect Dis 2012;54 Suppl 2:S109-S116.
17 Cardoso MR, Nascimento-Carvalho CM, Ferrero F, Alves FtM, Cousens SN. Adding fever to WHO criteria for diagnosing pneumonia enhances the ability to identify pneumonia cases among wheezing children. Arch Dis Child 2011;96:58-61.
18 Wingerter SL, Bachur RG, Monuteaux MC, Neuman MI. Application of the world health organization criteria to predict radiographic pneumonia in a US-based pediatric emergency department. Pediatr Infect Dis J 2012;31:561-564.
19 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352-359.
20 Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med 2011;18:1099-1104.
21 Shorr AF, Zilberberg MD, Reichley R, Kan J, Hoban A, Hoffman J, et al. Readmission following hospitalization for pneumonia: the impact of pneumonia type and its implication for hospitals. Clin Infect Dis 2013;57:362-367.
22 Aliberti S, Di Pasquale M, Zanaboni AM, Cosentini R, Brambilla AM, Seghezzi S, et al. Stratifying risk factors for multidrug-resistant pathogens in hospitalized patients coming from the community with pneumonia. Clin Infect Dis 2012;54:470-478.
23 Teka Z, Taye A, Gizaw Z. Analysis of risk factors for mortality of in-hospital pneumonia patients in Bushulo Major Health Center, Hawassa, Southern Ethiopia. Science J Public Health 2014;2:373-377.
24 Swarnkar V, Abeyratne U, Chang A, Amrulloh Y, Setyati A, Triasih R. Automatic identification of wet and dry cough in pediatric patients with respiratory diseases. Ann Biomed Eng 2013;41:1016-1028.
25 Kosasih K, Abeyratne UR, Swarnkar V, Triasih R. Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis. IEEE Trans Biomed Eng 2015;62:1185-1194.
26 Sainani KL. Logistic regression. PM R 2014;6:1157-1162.
27 The Acute Respiratory Ifection Diagnostic Aid (ARIDA) Project. http://www.unicef.org/innovation/innovation_81722.html (accessed December 22, 2016).
28 World's first pneumonia innovations summit unveils next generation prevention, diagnostic and treatment innovations. http://www.malariaconsortium.org/news-centre/worlds-first-pneumonia-innovations-summit-unveils-next-generation-prevention-diagnostic-and-treatment-innovations.htm (accessed November 12, 2015).
29 Abeyratne UR. Emerging tools for pneumonia diagnosis in resource poor regions. http://www.malariaconsortium.org/userfiles/file/Dr%20Abeyratne_slides.pdf (accessed January 27, 2017).
30 Abeyratne UR, Swarnkar V, Setyati A, Triasih R. Cough sound analysis can rapidly diagnose childhood pneumonia. Ann Biomed Eng 2013;41:2448-2462.
31 Floyd J, Wu L, Hay Burgess D, Izadnegahdar R, Mukanga D, Ghani AC. Evaluating the impact of pulse oximetry on childhood pneumonia mortality in resource-poor settings. Nature 2015;528:S53-S59.
32 Ginsburg AS, Delarosa J, Brunette W, Levari S, Sundt M, Larson C, et al. mPneumonia: development of an innovative mHealth application for diagnosing and treating childhood pneumonia and other childhood illnesses in low-resource settings. PLoS One 2015;10:e0139625.
33 Emdin CA, Mir F, Sultana S, Kazi A, Zaidi AMK, Dimitris MC, et al. Utility and feasibility of integrating pulse oximetry into the routine assessment of young infants at primary care clinics in Karachi, Pakistan: a cross-sectional study. BMC Pediatr 2015;15:1-11.
34 Mulholland EK, Simoes EA, Costales MO, McGrath EJ, Manalac EM, Gove S. Standardized diagnosis of pneumonia in developing countries. Pediatr Infect Dis J 1992;11:77-81.
35 Thaver D, Ali SA, Zaidi AK. Antimicrobial resistance among neonatal pathogens in developing countries. Pediatr Infect Dis J 2009;28:S19-S21.
36 Källander K, Nsungwa-Sabiiti J, Peterson S. Symptom overlap for malaria and pneumonia-policy implications for home management strategies. Acta Trop 2004;90:211-214.
37 Naydenova E, Tsanas A, Casals-Pascual C, De Vos M. Smart diagnostic algorithms for automated detection of childhood pneumonia in resource-constrained settings. IEEE Global Humanitarian Technology Conference, 2015: 377-384.

Received July 13, 2015
Accepted after revision March 17, 2016

(Supplementary information is linked to the online version of the paper on the *World Jounal of Pediatrics* website)