Using Bayesian networks to diagnose childhood illness in low- and middle-income countries: Case of malaria in Malawi

Sanya Bathla Taneja, Marek J. Druzdzel, Gerald P. Douglas

Introduction

Infectious diseases such as malaria and pneumonia are responsible for the majority of under-5 deaths in low- and middle-income countries. Early diagnosis and prevention can help in reducing the global burden of under-five morbidity and mortality. The Integrated Management of Childhood Illness (IMCI) protocol developed by the World Health Organization (WHO) is a set of clinical guidelines for the symptomatic management of childhood illnesses. Clinics for children under-5 are often characterized by limited availability of resources, unavailability of diagnostic testing facilities, and lack of expert clinicians.

Limitations of IMCI

- Diagnosis is a 'black box' as IMCI is designed for integrated case management of children under 5 years
- Symptom overlap makes differential diagnosis difficult using IMCI
- Limited support for non-malarial febrile illnesses leads to over-prescription of antimalarials and extensive use of malaria rapid diagnostic tests (mRDT)
- Minimal adherence to protocol due to various reasons such as laborious steps and high patient burden

Methods

We develop a Bayesian network for diagnosis of malaria in children under 5 years using clinical signs and symptoms.

Data

We use the Malawi Service Provision Assessment (SPA) conducted by the Ministry of Health in this study. The SPA survey data is publicly available and contains 3,441 observations of sick children aged 2-59 months. The dataset contains details of provider diagnosis, caretaker exit interview, and a limited re-examination by an expert healthcare provider for all children presenting at outpatient facilities in Malawi with an illness complaint.

Model Development and Validation

- Bayesian model is built using GeNIe where the clinical signs and symptoms of malaria in children, malaria diagnosis, along with the mRDT test results, form the nodes of the network. (Figure 1)
- An arc drawn from malaria to fever, with malaria as the parent node and fever as a child node, implies that malaria can cause fever.
- Conditional probability distributions for each node are learnt in the pre-defined network using the Expectation Maximization (EM) algorithm.
- The model is validated using 5-fold cross validation. We present the accuracy, area under ROC curve and contingency table for the diagnosis classification in the results.

Acknowledgements

The models described in this paper were created using the GeNIe Modeler, available free of charge for academic research and teaching use from BayesFusion, LLC, http://www.bayesfusion.com/.

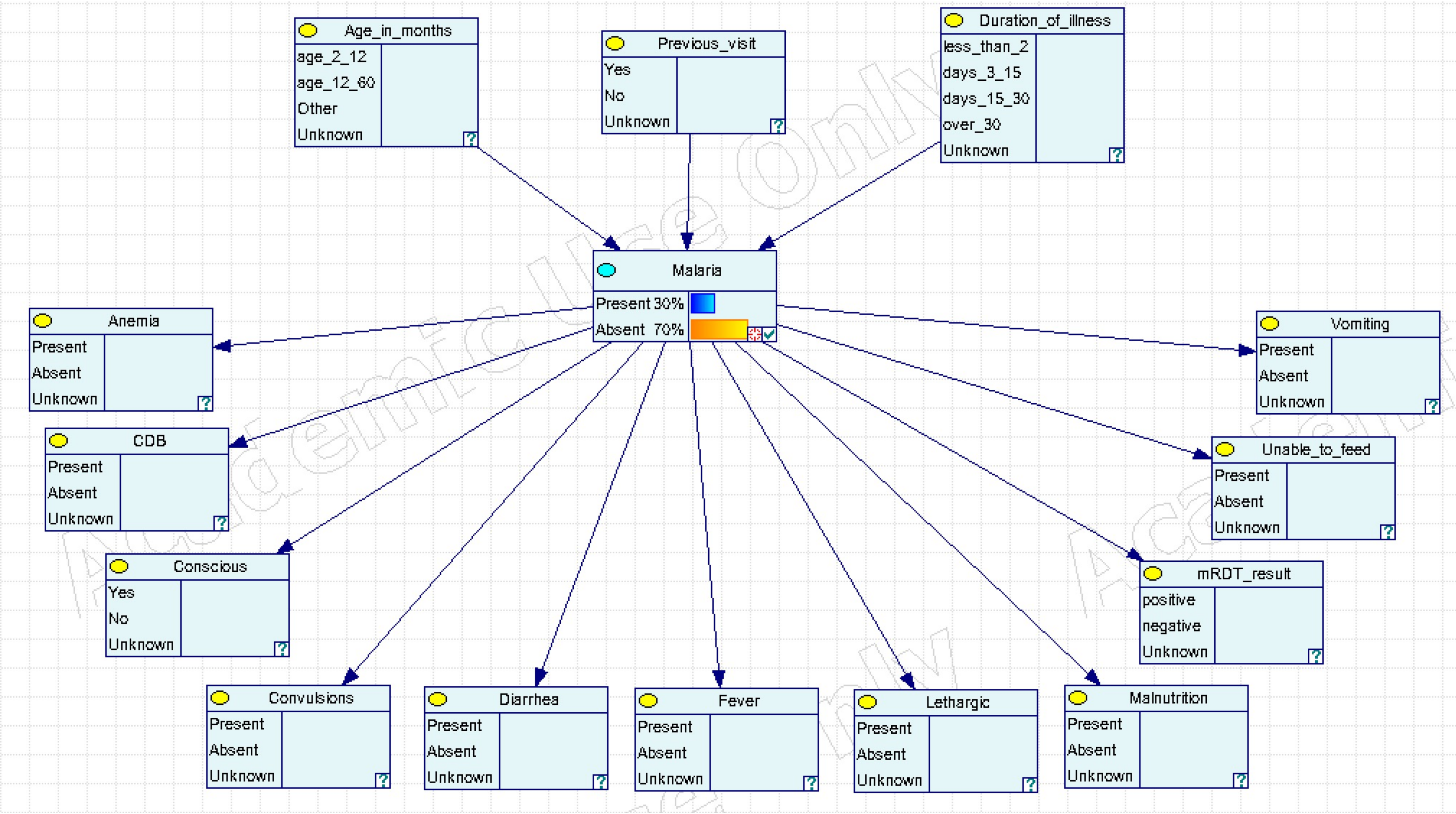
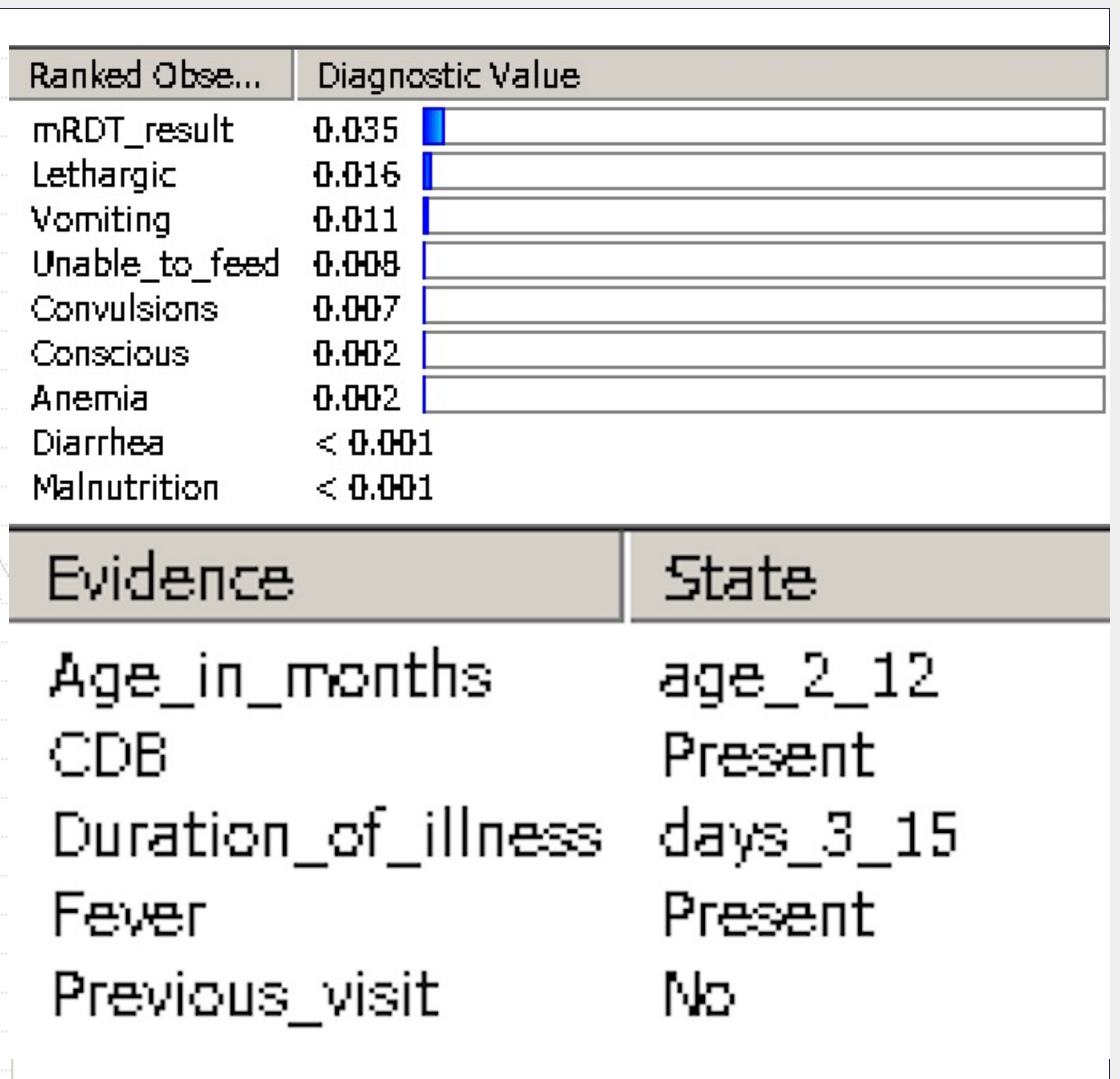


Figure 1: Malaria diagnostic model with variables as nodes and arcs representing causal relationships. When no other variables are observed, the probability of Malaria: Present is 0.3. Source: GeNIe



Ranked Obse...	Diagnostic Value
mRDT_result	0.035
Lethargic	0.016
Vomiting	0.011
Unable_to_feed	0.008
Convulsions	0.007
Conscious	0.002
Anemia	0.002
Diarrhea	< 0.001
Malnutrition	< 0.001

Evidence	State
Age_in_months	age_2_12
CDB	Present
Duration_of_illness	days_3_15
Fever	Present
Previous_visit	No

Figure 2: (a) Observed values in model and (b) Ranked list of signs and symptoms by diagnostic value. Source: GeNIe

Results

- Exploratory data analysis shows that the SPA data suffers from incompleteness, limited conclusive diagnosis fields, missing data entries and lack of validation.
- There are 3,433 observations excluding empty rows and rows with unknown malaria diagnosis. Out of 3,433 children, 988 (around 30%) present with positive malaria diagnosis.
- The Bayesian model correctly diagnosed 74.8% of the 3,433 combined cases in the test dataset.
- The average area under the ROC curve for the test sets is 0.74.
- Table 1 shows the contingency table for the classification results. The contingency table presents the number of malaria cases classified correctly given the total malaria cases present and absent.

	Malaria Present	Malaria Absent
Predicted Present	374	251
Predicted Absent	614	2194
Total	988	2445

Table 1: Contingency table for classification of malaria using Bayesian model

Selected References

1. Chart Booklet Integrated Management of Childhood Illness [Internet]. 2014 [cited 2019 Jul 28]. Available from: www.who.int
2. Lufesi NN, Andrew M, Aursnes I. Deficient supplies of drugs for life threatening diseases in an African community. BMC Health Serv Res. 2007;7:1–7.

Discussion and Future Work

For a sick child presenting with symptoms to a Health Surveillance Assistant (HSA), the Bayesian model presents the list of signs and symptoms to be observed that will likely lead to the most efficient diagnosis. This list is ranked according to the diagnostic value of each node in the network and provides crucial information regarding the next steps to be taken for diagnosis (Figure 2).

This leads to a more efficient diagnostic process wherein HSAs may rely on the model to guide the diagnosis. Moreover, the network parameters can be improved the longer the model is used by incorporating new evidence from sick child observations into the probabilities, increasing the accuracy of the model, and consequently providing better diagnosis to the patient.

Although this study presents a unique approach towards malaria diagnosis in under-5's, we faced several challenges in finding an appropriate dataset for the model. We have limited confidence in the current model as the data dictionary and coding of variables in the dataset was fairly confusing. The lack of confirmed laboratory diagnosis of the disease is the biggest obstacle in the SPA dataset. There is a call for more comprehensive disease datasets for children to develop models with better performance in the future.

- We plan to extend this study further to include other common childhood illnesses such as pneumonia. We also wish to explore the possibility of adding the HIV status of a child in the model as this is a known risk factor for infectious diseases.
- The current model has only been internally validated using the SPA dataset. For a more complete evaluation, the model must be deployed at the primary points of care. This will also allow us to improve the diagnostic model based on feedback from users.

University of Pittsburgh

Department of Biomedical Informatics