Quaderni di Statistica
Vol. 14, 2012

Bayesian Machine Learning Approaches for Longitudinal Latent Class Modelling to Define Wheezing Phenotypes to Elucidate Environmental Associates

Danielle Belgrave, Angela Simpson, Iain Buchan, Adnan Custovic
The University of Manchester
E-mail: danielle.belgrave@manchester.ac.uk, angela.simpson@manchester.ac.uk, buchan@manchester.ac.uk, adnan.custovic@manchester.ac.uk

Christopher Bishop
Microsoft Reseacrh Cambridge

E-mail: chistopher.bishop@microsoft.com

Summary: Accurate phenotypic definition of wheezing in childhood can lead to a greater understanding of the distinct physiological markers associated with different wheeze phenotypes. This paper looks at Bayesian machine learning approaches using Infer.NET to define wheeze phenotypes based on both parental questionnaires and General Practitioner data on patterns of asthma and wheeze consultation within the first eight years of life. We illustrate a taxonomy of longitudinal latent class item response models with varying modelling assumptions to determine wheeze phenotypes (latent classes) for homogenous groups of children.

Keywords: Longitudinal Latent Class Analysis, Bayesian Inference, Infer.NET.

1. Introduction

It has been widely recognised that asthma is a heterogeneous disease. Accurate phenotypic definition of wheezing in childhood can elucidate our understanding of this underlying complexity of asthma to identify distinct physiological markers associated with different wheeze phenotypes. We used longitudinal latent class modelling to identify subpopulations (classes) of children who differ in patterns of wheeze trajectories during childhood, based on both complete medical records and parental assessment of wheeze at different time points. We tested the validity of these classes by examining their relations with measures of lung physiology and atopy.

2. Methods

2.1. Data Description

Data are taken from the Manchester Asthma and Allergy Study (MAAS), an unselected, prospective population-based birth cohort study designed to determine early life factors for the development of asthma and allergic disease. Participants attended follow-up at ages 1, 3, 5 and 8 years. Validated questionnaires were interviewer-administered at each time-point to collect information on parentally-reported symptoms and physician-diagnosed asthma. Current wheeze was defined as parentally-reported wheeze in the past 12 months. Additionally, a trained paediatrician extracted data from electronic and paper-based primary care medical records, including wheeze/asthma diagnosis, and hospital admissions for asthma/wheeze during the first 8 years of life; we calculated the child's age in days for each event. We analysed data from questions assessing the presence of wheeze for 1185 children using these two complementary measures of wheeze.

2.2. Statistical Methods

We analysed all data using Bayesian machine learning approach using Infer.NET (research.microsoft.com/en-us/um/cambridge/projects/infernet/). The Bayesian machine learning method provides a unified framework for modelling and quantifying uncertainty - employing probabilistic modelling strategies based on defining priors in such a way that probabilities can be associated with unknown parameters. This allows us to incorporate and compare different modelling assumptions with a greater degree of flexibility. The three steps for defining a model in Infer.NET are: i) the definition of a probabilistic model; ii) the creation of an inference engine for performing inference; and iii) the execution of an inference query. We used Variational Message Passing (VMP) approximation for inference. VMP gives a factorised approximation for the model distribution.

We modelled a longitudinal latent class item response model to determine latent classes for homogenous groups of children. We analysed data from our two complementary sources: parentally-reported current wheeze at four follow-ups, and physician-confirmed wheeze recorded in medical records within each year from birth to age 8 years. We assumed that each child belongs to one of N latent classes, with the number and size of classes not known a priori. We assumed a discrete trajectory logistic regression model with a class dependent random intercept and slope for the dichotomous variable representing the answer to the question “Has the child wheezed within the given time period?”. This is represented in Equation (1) for child i at time t with wheeze assessed by rater r (physician or parent). Children were assigned to the latent class with the largest posterior probability given a uniform Dirichlet prior.

$$Logit\{Pr(y_{itr}) = 1|x, class_i = k\} = \beta_0+\beta_1 t+\beta_2[rater]+\beta_3[timexrater]+\xi_{class}+\xi_{class}t \tag{1}$$

We then extended this model by adding a quadratic term to the level-1 linear change trajectory model so that Equation (1) becomes:

$$Logit\{Pr(y_{itr}) = 1|x, class_i = k\} = \beta_0+\beta_1 t+\beta_2[rater]+\beta_3[timexrater]+\xi_{class}+\xi_{class}t+\xi_{class}t^2 \tag{2}$$

This model allows us to assess a higher order of complexity for the latent class trajectory.

We also considered a Hidden Markov Model which takes into account the sequential patterns of wheeze and correlations between observations that are close together. We used wheeze assessment from both physician and parental data to infer a multinomial latent variable for each child in order to assign children to their most probable class. We inferred time-dependent transition probabilities for the eight time points. Children belonging to the same latent class are similar with respect to their transition probabilities which are assumed to come from the same probability distributions, whose parameters are, however, unknown quantities to be estimated.

These modelling approaches with varying assumed number of classes were compared for goodness-of-fit using model evidence. The model evidence evaluates the probability of generating the data from a model whose parameters are sampled from the prior distribution while penalising the model according to its complexity. Classes identified in the optimal model were validated to see whether these patterns of wheeze represented a higher risk of various markers of asthma severity.

3. Results

Based on the model evidence, the optimal model was the discrete trajectory model with class dependent random Intercept and random linear slope with five latent classes of wheeze. Based on our interpretation of the results, we have labelled these classes as “No Wheeze” (53.3%), “Transient Early Wheeze” (13.7%), “Late-onset Wheeze”(16.7%), “Persistent Troublesome Wheeze” (3.2%) and “Persistent Controlled Wheeze” (13.1%) (Figure 1). Mixed effects models revealed significant differences in these 5 wheeze phenotypes.“Persistent Troublesome Wheeze”, “Late-onset Wheeze” and “Persistent Controlled Wheeze” were found to be associated with atopy ($p<0.001$). There was a significant difference in initial Specific Airway Resistance (sRaw) values for the five classes

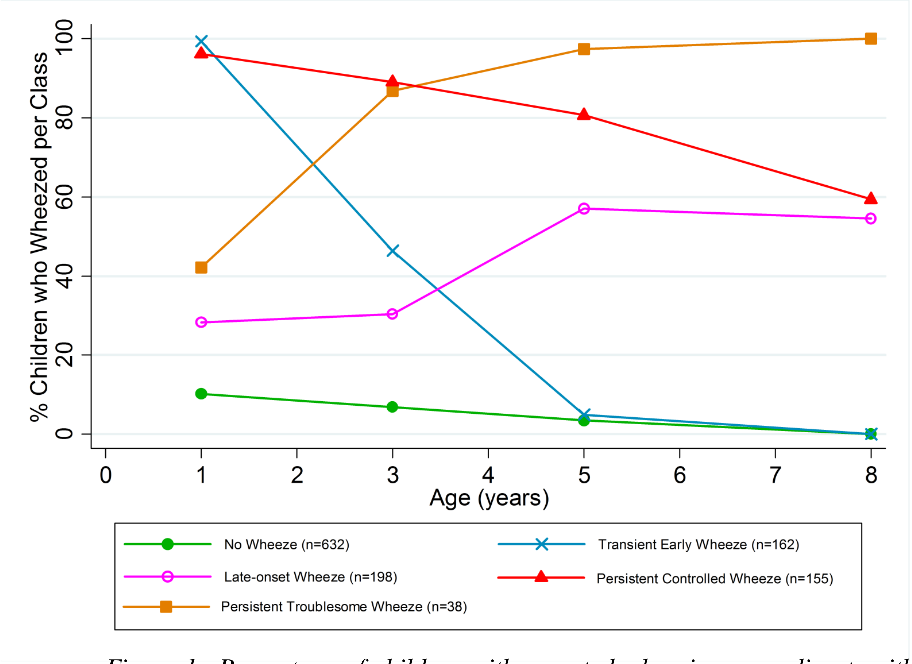

Figure 1. Percentage of children with reported wheezing according to either parental and physician reporting. The number of children in each class is denoted in parentheses

(p<0.001). Children with early onset wheeze and persistent controlled wheeze show significant increase in sRaw over time (p=0.03 and 0.02 respectively).

4. Conclusion

The joint modelling of observations from the general practitioner and parental reporting of childhood wheeze enables us to identify phenotypes of wheeze with greater accuracy and determine their risk factors and characteristics.

References

Bishop C.M. (2006), Pattern Recognition and Machine Learning, *Springer*

Minka T., Winn J.M., Guiver J.P., Knowles D.A. (2010): Infer.NET 2.4 *Microsoft Research Cambridge. http://research.microsoft.com/infernet.* (2nd ed.)