

A Comparison of Frequentist and Bayesian Approaches to Latent Class Modelling of Susceptibility to Asthma and Patterns of Antibiotic Prescriptions in Early Life

Student Poster Presentation

Danielle Belgrave*¹, Christopher Bishop², Adnan Custovic³, Angela Simpson³, Aida Semic-Jusufagic³, Andrew Pickles⁴, Iain Buchan¹

¹ North-West Institute for Bio-Health Informatics, The University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL.

² Microsoft Research Cambridge, Roger Needham Building, 7 JJ Thomson Avenue, Cambridge, United Kingdom, CB3 0FB

³ Respiratory Department, The University of Manchester, University Hospital South Manchester, Southmoor Road, United Kingdom, M23 9LT

⁴ Institute of Psychiatry, Kings' College London, Box P, De Crespigny Park, London, United Kingdom, SE5 8AF

*Corresponding author: danielle.belgrave@manchester.ac.uk

Abstract: The assessment of patterns of antibiotic use in early life may have major implications for understanding the development of asthma. This paper compares a classical generalized latent variable modelling framework and a Bayesian machine learning approach to define latent classes of susceptibility to asthma based on patterns of antibiotic use in early life. We compare the potential advantages of each method for elucidating clinically meaningful phenotypes or classes.

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

1 Introduction

The assessment of patterns of antibiotic use in early life may have major implications for our understanding of the development of asthma. Within the medical literature, antibiotic use has been investigated as having a causal association with asthma. We hypothesise that antibiotic use in early life, rather than being causally related to asthma gives an indication of a child's susceptibility to infection with a heightened response to exposure since such children are more likely to receive antibiotics early on in life due to their immunodeficiency. Thus antibiotic use in early life can be

used as a marker in order to identify children who are more susceptible to outcomes of exacerbation of wheeze and asthma symptoms. The aim of this project is to identify latent classes of susceptibility to characterize children according to susceptibility based on patterns of early-life antibiotic use and investigate whether this latent class is predictive of contemporaneous and future asthma and wheeze symptoms. We infer that antibiotic use picks up a signal of something that occurs very early in life, and which is completed by 24 months of age.

2 Methods

The Manchester Asthma and Allergy Study (MAAS) is an unselected, prospective population-based birth cohort study designed to determine early life factors for the development of asthma and allergic disease. Subjects were recruited prenatally and followed prospectively. A trained physician extracted the information on antibiotic prescription receipt and symptoms of asthma/wheezing from the primary care medical records (n=916). Based on a longitudinal model for antibiotic use within the first 2 years of life, latent class analysis was carried out to obtain a phenotypic definition of susceptibility. We then investigate whether these latent classes of susceptibility are predictive of contemporaneous and future asthma and wheeze symptoms. We describe and use two different statistical approaches for defining latent classes of susceptibility to asthma: a classical generalized linear latent and mixed models framework using the `gllamm` package in STATA and a Bayesian machine learning approach using `Infer.NET`.

Using the classical approach to latent class analysis we formulated a longitudinal trajectory model which allows us to hypothesize that there may be subgroups of children who, because of differing maturity of their immune response, have changing levels of susceptibility over time. We specified this as a two-level random-coefficient logistic regression model for antibiotic use with level-1 units as the monthly measurement occasions and the level-2 units as children. This model characterises the child's susceptibility through their age and exposure to two particular known risk factors (older siblings and day-care) and by membership of different possible classes defined by the intercept and slope in the regression equation for antibiotic use y_{ij} of child i at time j which was specified as:

$$\text{Logit}\{Pr(y_{ij} = 1|x_{ij}, c_i = k)\} = \beta_{0k} + \beta_{1k}x_{1i} + \beta_{2k}x_{2ij} + \beta_{3k}x_{3ij} \quad (1)$$

where x_{1i} is the time point for a specified child i . x_1 represents monthly time periods and $x_1 = 1, \dots, 24$ months; x_{2ij} is child i 's day-care attendance at time j ; and x_{3i} is the number of older siblings child i has which

remains constant at all time points j . We also introduce a prior distribution $Pr(c_i = k)$ over the classes given by a multinomial distribution.

We assume that each child belongs to one of a set of N latent classes, with the number of classes and their size not known a priori. Other than random temporal fluctuation, each child's pattern of antibiotic prescription is to be explained by their belonging to a particular class of susceptibility. Children belonging to the same class are similar with respect to the observed variables in the sense that their observed scores are assumed to come from the same probability distributions, whose parameters are, however, unknown quantities to be estimated. Using empirical Bayes techniques, children are assigned to the latent class with the largest posterior probability. We also consider a restricted random-intercept form of this model in which the classes are allowed to differ in their intercepts, but not in their slopes i.e. in which the relative susceptibility remained constant over time. These models were fit using `gllamm`, a program implemented in Stata (www.stata.com) to fit generalized linear latent and mixed models.

We then investigated parallel models using a machine learning approach with the Bayesian inference software Infer.NET. 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. 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. Since the Bayesian approach to statistical modelling enables us to quantify model uncertainty through the incorporation of priors, we assumed uninformative priors for y_{ij} . Variables x_{1i} , x_{2ij} and x_{3ij} are specified as a vector array X with a vector of coefficients β . The k unobserved latent classes are accompanied by a random temporal fluctuation or noise, ξ_k and c_i is assumed to be multinomial over k classes with a prior uniform Dirichlet distribution (Dirichlet (1,1)) and the random noise has a prior Gaussian distribution(0,1).

We compare models that assume varying numbers of latent classes using the Bayesian Information Criterion as a measure of goodness-of-fit. We then investigate whether the inferred phenotypes of susceptibility predict current or future asthma and wheeze symptoms using conventional time-to-event analyses.

3 Results

Using the classical generalized linear latent and mixed models framework, we identified a model with three distinct latent classes of susceptibility based on patterns of antibiotic use in the first two years of life. Based on our interpretation of the model, Class 1 were children resilient to infection (31.1%), Class 2 showed a normal immune response (55.7%) and Class 3 were susceptible to infection (13.2%). Compared to Class 1 and Class 2, children in Class 3 had a significantly higher hazard of reported asthma or wheeze symptoms in the first three years of life (HR=3.72 [95% CI 2.72 – 5.10, $p < 0.01$] and 1.61 [95% CI 1.25 – 2.09, $p < 0.01$] respectively. Class 2 had a greater hazard of experiencing exacerbations of asthma and wheeze symptoms than Class 1 (HR=1.90 [95% CI 1.21 – 2.98, $p < 0.01$]) however, after the third year of life, this hazard ratio ceased to be statistically significant (HR=1.39 [95% CI 0.79 – 2.45, $p = 0.25$]). Similar results were obtained using a Bayesian machine learning framework. We demonstrate the potential advantages of Bayesian models for elucidating clinically meaningful phenotypes.

4 Conclusion

By analysing trajectories of antibiotic use in early life, we identified a group of children with high susceptibility to the development of asthma. Our results suggest that antibiotic use in early life indicates a child's susceptibility to infections. Since the Bayesian and frequentist approaches provided different perspectives for identifying the latent classes, with concordant results, the combination of methodologies was complementary – Bayesian extensions to classical epidemiology have the potential to shape hypotheses with more complete use of the data and current knowledge.

References

- Bishop, C.M (2006). *Pattern Recognition and Machine Learning*. USA: Springer.
- Custovic, A., Simpson, B.M., Murray, C.S., Lowe, L., and Woodcock, A. (1994). The National Asthma Campaign Manchester Asthma and Allergy Study. *Paediatric Allergy Immunology, Supplement.*, **15**, 32-7.
- Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – A modified latent structure approach *American Journal of Sociology*, **79**, 1179-1259
- Rabe-Heskett, S., Skrondal, A., Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, **128**, 301-23.