

Statement of Scholarly Interest for Ann Michelle Norris

My current research involves developing new models for diagnostic screening data. Diagnostic screening data refer to data resulting from any test used to determine whether or not a subject has a certain disease or infection. A variety of screening tests, yielding different types of data, are currently in use. These include serology tests which are designed to measure an antibody reaction, microscopic examination which involves looking for physical evidence of disease in a lab culture or tissue sample, and Polymerase Chain Reaction (PCR) tests which are based on locating DNA that match a particular infectious agent. Data from these tests are usually continuous, discrete and ordinal, respectively.

Since many of the tests in common use are imperfect, misclassifying both diseased and non-diseased subjects, scientists are often interested in estimating the sensitivity and specificity of a screening test. The sensitivity is the probability a diseased subject is correctly classified whereas specificity is the probability a non-diseased subject is correctly classified. Even if the raw test results are continuous or ordinal, they are typically reduced to binary classification with scores above a certain cutoff value classified as diseased and those below as non-diseased. Estimation of a test's sensitivity and specificity is sometimes complicated by the fact that the true disease status of the subjects is unknown because a so-called gold standard, or perfect, test is either not available, too costly or too invasive. Testing in the no gold standard case is currently an active area of research, and models have been developed to allow estimation of sensitivity and specificity for a variety of scenarios involving no gold standard data. Historically, these models were developed for binary cross-sectional data.

Recent work has extended these models to screening tests with continuous outcomes and to longitudinal screening tests (where repeated measures are taken on each subject). Models for continuous data often require that the data for diseased and non-diseased subjects be transformable to separate normal distributions, which requires some ingenuity in the no gold standard case. Longitudinal models research has focused on binary outcomes.

My primary research involves constructing models for longitudinal diagnostic screening data in the no gold standard case. Moreover, I extend current models by jointly modeling continuous serology and binary microscopic examination outcomes and including a change-point corresponding to infection time. The proposed models are mixed effects models which incorporate random coefficients for a parametric mean function in the continuous part of the model. Random effects are usually assumed to be normally distributed; however, this assumption may not be tenable for screening data due to clustering of subjects who have similar immune responses. I incorporate nonparametric modeling of random effects to allow more flexibility. Nonparametric models do not constrain the distribution of a variable to any specific form such as the normal, gamma or log normal, for example. They do not even enforce symmetry; rather they allow the data to drive the shape of the distribution. Because I make inference within the Bayesian framework, prior distributions on the space of random effects

distributions must be specified. I use the popular Dirichlet Process mixture prior for this purpose and hope to extend this work to other priors such as the Mixture of Finite Polya Trees prior.

The Bayesian framework is particularly attractive for this sort of modeling because it allows for the input of expert information on screening test performance and other biological aspects of the problem in a probabilistically coherent manner through the specification of appropriate prior distributions. Because the models I develop are sufficiently complex to defy analytic solution, I use Markov Chain Monte Carlo (MCMC) computational methods to estimate parameters of interest. MCMC methods exploit modern computing power to draw large samples from high-dimensional posterior distributions which are of inferential interest. Prior to important developments in MCMC sampling by Gelfand and Smith in 1990, complex Bayesian models such as this one were intractable. The proposed model is additionally complicated by the fact that the dimension of the parameter space changes depending on whether or not the change-point (or infection) has occurred. Models with varying-dimensional parameter spaces require special MCMC techniques such as Green's Reversible Jump MCMC. As an alternative, I develop a novel parameterization and a corresponding Metropolis-Hasting sampler which circumvent the need for Reversible Jump MCMC. It is hoped that, by incorporating all available data through priors and longitudinal modeling, these models and estimation methods will increase the accuracy with which scientists diagnose diseases in humans and animals.

In addition to diagnostic screening, longitudinal data, nonparametric modeling and MCMC methods, I have interests in several other areas of research. I had the opportunity to work as a research assistant on projects involving gene sequence data. In particular, I worked on using gene sequence data to classify foot-and-mouth disease virus isolates by location and host animal, using functional data from laboratory tests on foot-and-mouth disease isolates to identify the strain of the virus and constructing phylogenetic trees to infer the evolutionary relationships among a group of isolates. The science involved in bioinformatics is fascinating, and I would enjoy further collaborative research in this field. Further back in my past, I worked for seven years as a sampling statistician for the California Department of Conservation and, from this experience, gained an interest in public and environmental policy research. Finally, I would be very interested in research in mathematics and statistics education, particularly in assessing the role and effectiveness of technology in teaching mathematics, developing new technology tools for teaching mathematics and in determining the best methods for developing problem-solving and critical thinking skills in mathematics students.