**Introduction**

In biomedical studies, investigators are often interested in predicting future or missing observations of subjects based on their historical data, referred to as dynamic prediction. Traditionally, mixed models or joint modeling have been used for these problems. However, these methods are often limited in flexibility when dealing with densely measured functional data. In addition, out-of-sample prediction is challenging under model frameworks above, especially with non-Gaussian outcomes (e.g., binary and count data), because the estimation of individual random effect can be computationally intensive. To address these problems, we propose a novel method for dynamic prediction based on functional data analysis methods, which allows for more flexible correlation structure, can be generalized to non-Gaussian outcomes and is computationally feasible.

**Method**

The proposed new method is an extension of Generalized Functional Principal Component Analysis (FPCA) and Functional Mixed Models. Assume that the generalized functional outcome can be characterized by a latent Gaussian function. Specifically, the outcome at a specific time follows an exponential family distribution parameterized by the value of latent function at that time. Such latent functions can be approximated with the linear combination of a set of basis functions, which naturally lead us to dimension reduction methods such as FPCA. Here, each latent function can be characterized with a set of eigenfunctions and their corresponding subject-specific coefficients, often referred to as PC scores. Since eigenfunctions does not change across population, any point in the functional domain can then be estimated with these subject-specific scores.

Practically, the prediction procedure above consists of the following steps: 1) Binning the data across functional domain into small, equal-length, non-overlapping intervals; 2) Fitting local generalized mixed models (GLMM) at every bin to obtain an estimated latent function for each subject on the binned grid; 3) Use FPCA to smooth the estimated latent functions, also obtain eigenfunctions and estimates of variance components for prediction. With this information from FPCA model, we can obtain estimates of subject-specific PC scores based on partial observations, thus recover the unobserved, latent Gaussian function on the binned functional domain.

The recovery of latent Gaussian functions includes two different cases: in-sample estimation and out-of-sample prediction. The former refers to situations when the full functional tracks are observed and used for model fitting. The individual functions are then “in-sample” subjects. The latter, on the other hand, refers to new function tracks that have not been included in steps 1-3, namely “out-of-sample” subjects. If the out-of-sample function tracks are incomplete, we will be able to recover their full latent function tracks with information obtained from step 3. This procedure is called out-of-sample prediction and is more in line with the “dynamic prediction” procedure in practice.

**Results**

We have designed and implemented a simulation study with 500 subjects, with binary outcomes generated from a latent Gaussian function with four eigenfunctions. We have also compared the predictive performance of our model to Generalized Linear Mixed Models using Adaptive Gaussian Quadrature (GLMMadaptive), which is a popular and computationally feasible approach for dynamic prediction of longitudinal non-Gaussian data. The prediction of latent Gaussian process was evaluated with Integrated Squared Error (ISE), defined as the sample average of total squared error over unobserved bins for each subject. The prediction of binary outcome function was evaluated by the average Area under the Receiver Operating Characteristic (ROC) curve (AUC), where ROC curve for each subject is built on the probabilistic predictions over the unobserved time points on the original grid. While probabilistic prediction for the proposed method is only available at bin midpoints, the values between two bin midpoints are estimated using linear interpolation.

The preliminary results revealed that our proposed method achieved significantly better out-of-sample predictive performance on both latent Gaussian function and binary outcome function. The discrepancy between ISE and AUC of the two methods also increased with longer observed tracks. The proposed method was also much more time efficient. While the model fitting process was much faster, computation time spent on prediction given the fitted models is also much shorter than the GLMMadaptive model.

**Discussion**

Simulation results above revealed that the proposed method can achieve better predictive performance with much smaller computational burden with dense non-Gaussian repeated measures. In fact, few existing methods are feasible for dynamic prediction on the scale of the simulated datasets. The proposed method can fill in this gap and has the potential to be widely applicable to large datasets in a variety of fields, such as accelerometer data, longitudinal imaging data, etc.

However, we are not clear about the effect of arbitrary decisions in the binning and local GLMM steps on the final predictive performance, such as the effect of bin width, number of observations in each bin, and the model form of GLMMs. While some of them can be treated as hyperparameters and cross-validated, others have more complicated indications. In addition,

the procedure above allowed us to make predictions on the binned grid, but not the original grid. One potential way to extend prediction back to the original grid is through interpolation. Since the bins are set up to me small so that binned observations points are also dense, interpolation should perform well filling in function values between two discrete points.

The out-of-sample prediction interval is another challenging issue that needs further exploration. Just like random effects of generalized mixed models, there is no closed-form solution for the estimates of PC scores of generalized functions, neither for their covariance. A general strategy to address this issue is to use resampling procedures, such as bootstrap, to estimate its distribution. We may also be able to get conditional variance estimates using observed Fisher information. However with the additional constraints, such estimates can be biased and not interpretable.