**Design of the Nephrotic Syndrome Study Network (NEPTUNE):**

**A Multi-Disciplinary Approach to Understanding Primary Glomerular Nephropathy**

Statistical Methodology Supplement

The primary analysis plan involves deriving a prediction model to rigorously evaluate and identify potential predictors for study endpoints. Subsequently, the overall prediction model will be used to assess association via odds ratios for targeted sub-group analysis. For example, specific phenotypic and molecular information can be utilized to determine the probability of achieving remission which will be helpful for prognosis and developing therapeutic strategy.

For biomarker discovery and validation, we will randomly split the study sample into the Training and the Test sets at the ratio of $\frac{2}{3}$n :$\frac{1}{3}n$[1, 2]. The former is used to build a prediction model and the latter is for internal validation. Selection of biomarkers (e.g. gene expression levels, or proteomic markers) will proceed in two steps. Initially, we will perform univariate screening based on p-values obtained from tests of association between a study endpoint and a biomarker. This step will enable us to determine a relatively small pool of promising biomarkers. The cut-off by a pre-specified false discovery rate or a fixed number of top predictors may be used to determine the pool size. It is known that univariate screening may produce many false signals [3]; for example, a biomarker that is not associated with disease outcome but strongly correlated with a predictive biomarker could be selected. To mitigate this risk, an additional step of joint screening for those selected biomarkers in the pool will be performed to refine the identified candidate biomarkers and remove false signals using the training set. Univariate screening will be performed using logistic regression models. Due to the potential for a large number of candidate biomarkers and clinical markers, joint screening will be carried out using LASSO [4] or LARS [5] regularized regression approach. In the joint LASSO based screening (e.g. Lemley et al.[6]) cross-validation will be used to determine tuning parameters and to calibrate the prediction model in the Training set. As part of calibration in the building of prediction model, we will assess the variable

selection stability in the LASSO-based joint screening through bootstrap samples generated from the training data [7].

The NEPTUNE Test dataset which is an independent NEPTUNE subcohort separated from the training dataset, will be used to generate the ROC curve to develop the final model based on the highest predictive power. Alternatively, Net Reclassification Index [8] may be used to quantify the change of prediction accuracy using the NEPTUNE Test cohort when choosing specific predictors to be included in the final model. In addition, model calibration will be examined. The procedure of choosing the tuning parameters in LASSO according to the smallest Akaike information criterion (AIC), the Bayes Information criterion (BIC) or cross-validation error provides one approach for model calibration. Other calibration measures, such as diagnostic residual plots and Hosmer-Lemeshow test for the goodness-of-fit in logistic regression model, will be considered.

Analysis of repeated measurements (e.g. estimated GFR) will be undertaken by the means of standard mixed-effects models supplemented by generalized estimating equations (GEE) models, for both biomarker screening and prediction model building. The resulting prediction model will enable us to use selected biomarkers and auxiliary clinical parameters to discriminate and predict patient's longitudinal trajectories [9].

To address the challenge in building the prediction model for combining molecular biomarkers and clinical variables, we propose to use multiple indices [10] which will allow dimension reduction among multiple predictors from various sources. Statistical power will be improved in such a model formulation because proper scaling on grouped predictors can be applied to reconcile differences of variables corresponding to patient characteristics. In addition, the synthesis of different sources of predictors provides an opportunity to further refine the selection of important predictors by removing redundant predictors.

In a longitudinal cohort with the complex data collection implemented in NEPTUNE, missing data may be present. In the study design, we assume 10% attrition rate for the longitudinal cohorts. For participant dropout that is independent of the underlying disease mechanism (i.e. missing completely at random), a valid approach is to analyze all available data using linear model or linear mixed-effects model [11]. In the case that the mechanism of

dropout is suspected to be dependent on observed covariates, likelihood estimation and inference in the mixed-effects models are still valid for all available data analysis. We will model the probability of missingness through a logistic model, and then incorporate estimated probability of missingness into the GEE estimation. Every attempt will be made to identify missing covariates. However, in the event that this data is unobtainable, we will invoke proper statistical strategies [12] to deal with this challenge.

Predictors, including biomarkers and clinical parameters, established in the NEPTUNE cohort will require additional validation with external cohorts. To this end, the NEPTUNE protocol has been shared with multiple international networks to enable future cross validation.

REFERENCES


1. Dudoit S FJ, Speed TP. Comparison of discrimination methods for tumor classification based on microarray data. *JASA* 2002; **97:** 77-87.

2. Ma S, Song X, Huang J. Regularized binormal ROC method in disease classification using microarray data. *BMC bioinformatics* 2006; **7:** 253.

3. Fan J, Lv, J. Sure independence screening for ultra-high dimenstional feature space. *J Roy Stat Soc B* 2008; **70:** 849-911.

4. Tibshirani R. Regularization shrinkage and selection via the lasso. *J Roy Stat Soc B* 1996; **58:** 267-288.

5. Efron B HT, Johnstone I, and Tibshirani R. Least angle regression. *Ann Stat* 2004; **32:** 407-499.

6. Lemley KV, Lafayette RA, Derby G*, et al.* Prediction of early progression in recently diagnosed IgA nephropathy. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* 2008; **23:** 213-222.

7. Wang S, Nan B, Rosset S*, et al.* Random lasso. *Annals of Applied Stat* 2011; **5:** 468-485.

8. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr*., et al.* Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine* 2008; **27:** 157-172; discussion 207-112.

9. Li Y, Wang, S., Song P.X.K., Wang, N., Zhou, J. Doubly regularized estimation and selection i linear and mixed-effects models for high dimensional longituindal data. *Journal of the American Statistical Association* 2012; **in revision**.

10. Xia Y. A multiple-index model and dimension reduction. *Journal of the American Statistical Association* 2008; **103:** 1631-1640.

11. Diggle P, Heagerty P, Liang, K-Y, and Zeger SL. *Analysis of longitudinal data*, 2nd edn. Oxford University Press: New York, 2002.

12. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Wiley: Hoboken, N.J., 2004.