Identification of clinical and genetic predictors of Parkinson's disease progression via Bayesian machine learning

GNS HEALTHCARE
DISCOVERING WHAT WORKS, AND FOR WHOM.

Jeanne C. Latourelle[1], Michael T. Beste[1], Tiffany C. Hadzi[1], Robert E. Miller[1], Jacob N. Oppenheim[1], Matthew P. Valko[1], Diane M. Wuest[1], Iya G. Khalil[1], Boris Hayete[1], Charles S. Venuto[2]

[1] GNS Healthcare, Cambridge, MA [2] Center for Human Experimental Therapeutics and the Department of Neurology, University of Rochester, Rochester, NY

Introduction

The clinical progression of Parkinson's disease (PD) is highly heterogeneous across patients. Identifying features predictive of the rate of disease progression can:

- provide insight into the mechanisms of disease process
- inform clinical trial enrollment
- aid clinical disease management

The aim of this study was to develop data-driven models of PD progression, separately for both motor and cognitive symptoms.

Our novel machine learning platform allows the identification of an optimal ensemble of multivariate predictors from a complex data set including a variety of clinical, genetic, molecular and imaging data.

Methods

Source Data and Study Population

Discovery Set: Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org)[1]. Participants with 2+years of follow-up available as of 12/28/15 were included

- 317 untreated PD patients identified within two years of diagnosis
- 118 age- and sex- matched healthy controls

Validation Set: 317 independent *de novo* PD subjects followed 7+ years from the Longitudinal and Biomarker Study in PD (LABS-PD)[2]

Modeling Approach and REFS™ Analytical Platform

Rate of clinical progression of two clinical domains, Motor and Cognitive, were estimated using linear mixed effects models of subject-specific annualized rate of change of the appropriate clinical assessment

- Motor: Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Parts II and III
- Cognitive: Montreal Cognitive Assessment (MoCA)

Potential predictors included medical evaluations (N=18), neurological imaging (N=8), genotyping (N=17,456), and CSF biomarkers (N=7).

GNS Healthcare's proprietary machine learning platform, Reverse Engineering and Forward Simulation (REFS™)[3] was used to build prediction models. Selection of a single model underestimates prediction error, thus REFS learns an ensemble of the most probable models (N=128) given the data (Figure 1).

- Ensemble constructed via Monte Carlo sampling of the posterior model landscape.
- Model additions/subtractions scored based on a maximum entropy structural prior with complexity also penalized by the Bayesian Information Criterion[4].
- Linear, additive, quadratic, and cubic terms allowed in order to accommodate non-linear effects and sub-populations.
- Confidence of a given relationship X→Y determined by frequency among ensemble.

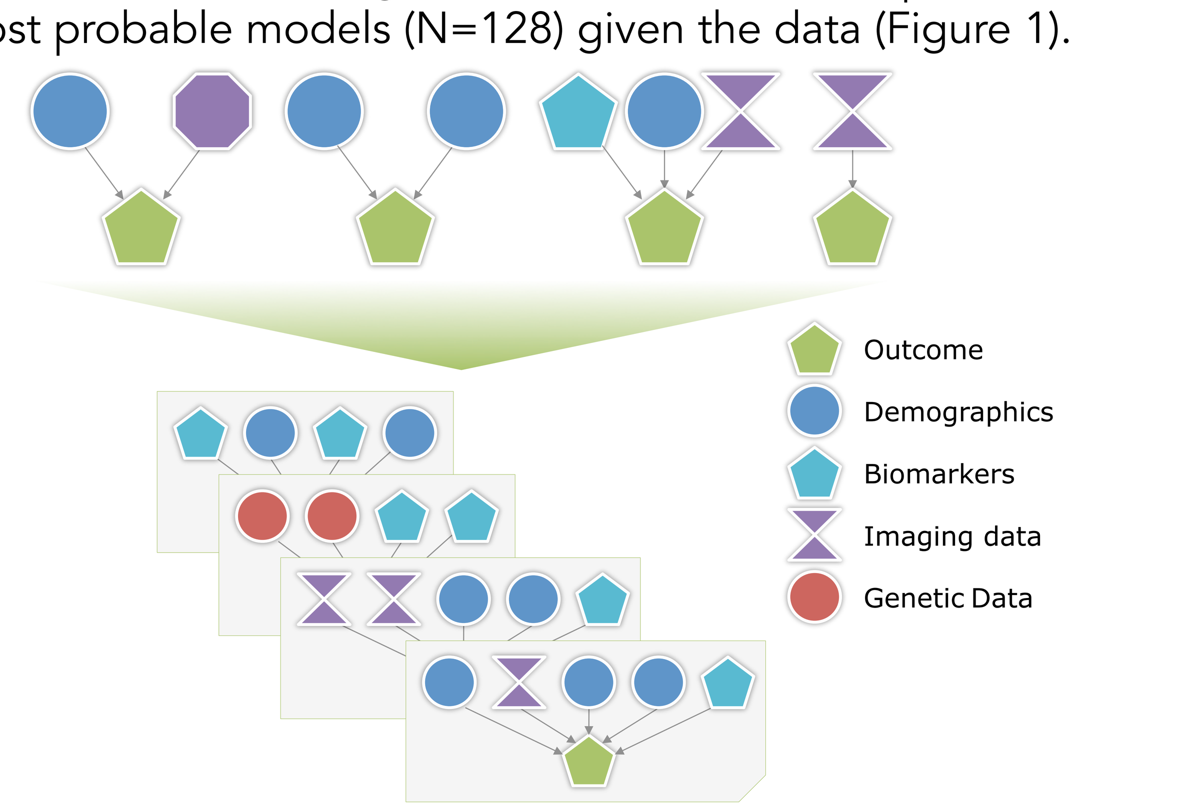

Figure 1. Visualization of REFS™ enumeration of model fragments and reverse-engineering of prediction model ensemble.

Results

Validation and Extensions of Prediction Models

Predictive performance estimated via 5-fold cross-validation of PPMI samples and LABS-PD samples, using Pearson R^2 for predicted vs. observed progression rates (see table)

Motor Progression Rates (MDS-UPDRS Part II & III units/year):

- Accuracy greater in cases than controls
- Reduced accuracy in untreated cases
- Not significant for later-stage cases

Cognitive Progression Rates (MoCA units/year):

- Accuracy greatest among untreated cases
- Less variability in accuracy across strata than in motor progression

Strata	Motor Progression				Cognitive Progression			
	PPMI		LABS-PD		PPMI		LABS-PD	
	N	R^2	N	R^2	N	R^2	N	R^2
All	639	0.41	317	0.09	473	0.48	317	0.17
Cases	522	0.27	317	0.09	356	0.48	317	0.17
Controls	117	0.01^{ns}	-	-	117	0.35	-	-
Untreated	296	0.19	27	0.15	135	0.55	27	0.14
Treated	226	0.05	290	0.11	221	0.45	290	0.20
Early stage	500	0.28	23	0.02^{ns}	342	0.50	23	0.31
Later stage	22	0.12^{ns}	294	0.11	14	0.04^{ns}	294	0.15

Validation and Extensions of Prediction Models

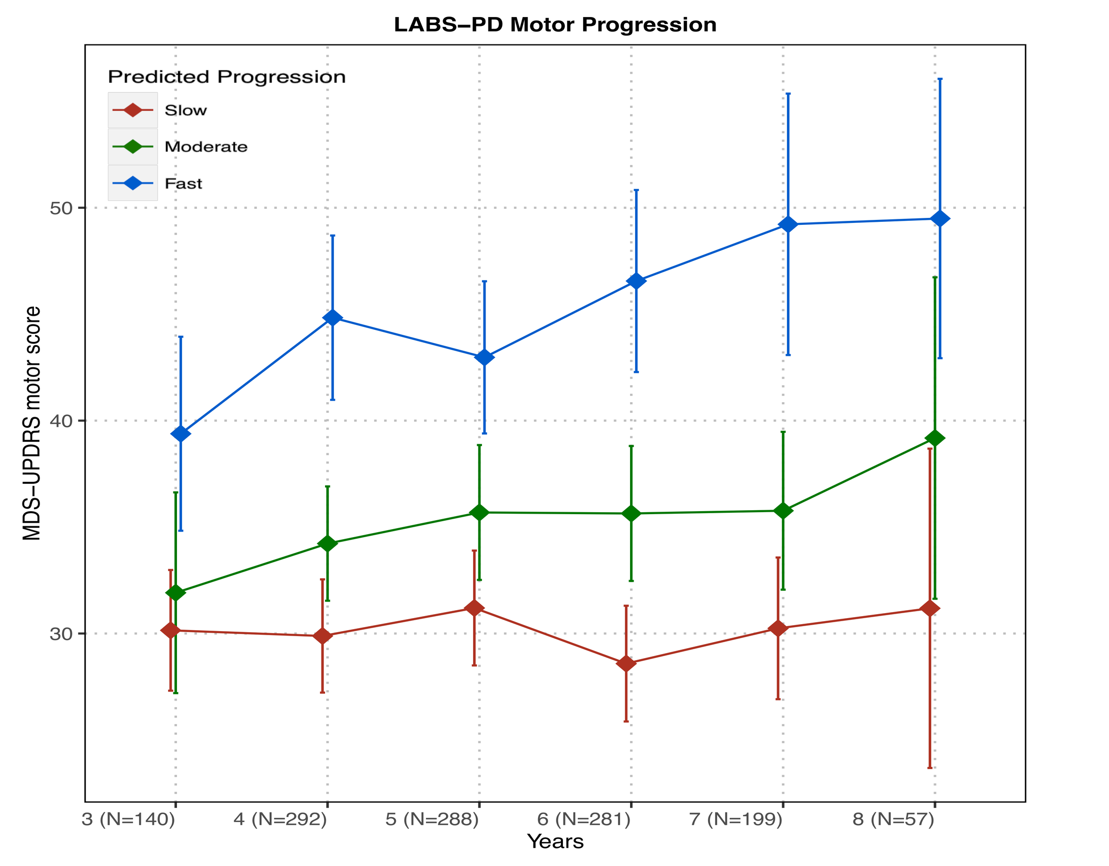

Figure 2. LABS-PD Median and 95% CI of motor scores over follow-up by model-predicted baseline progression categories.

In the independent LABS-PD cohort, the model ensembles demonstrate unequivocal ability to identify, early on, patients whose condition would deteriorate most rapidly (motor progression shown in Figure 2; cognitive progression not shown)

- "Slow", "Moderate", and "Fast" groups were defined by tertile splits of the individual calculated rates
- Significant differences between the slowest and fastest groups across all time points
- Moderate group shows significant difference from the fast group for all but the final year
- Slow and moderate groups were not as strongly differentiated

Ensemble Frequencies and Replication of Candidate Progression Biomarkers

Both ensembles combined both novel and established markers of disease progression.

Motor Progression Rates (MDS-UPDRS Parts II & III units/year):

- 80 unique predictors, 11 with >5% frequency
- Baseline motor score, PD status, SWEDD status, PD med use, and sex selected in over 90% of models

Cognitive Progression Rates (MoCA units/year):

- 205 unique predictors, 21 with >5% frequency
- Baseline age, MoCA score, CSF t-tau/Aβ1-42 ratio, PD med use, sex, African ancestry, motor score, PD status, and education selected in over 90% of models
- Caudate/Putamen count density ratio selected in 70%; higher ratios predict slower decline

Genetic variants were among the selected features, but at lower frequencies, and often in interactions.

The most common genetic predictor of motor progression, a novel interaction between rs9298897 (intronic *LINGO2*) and rs17710829 (2q14.1) was replicated in LABS-PD (Figure 3).

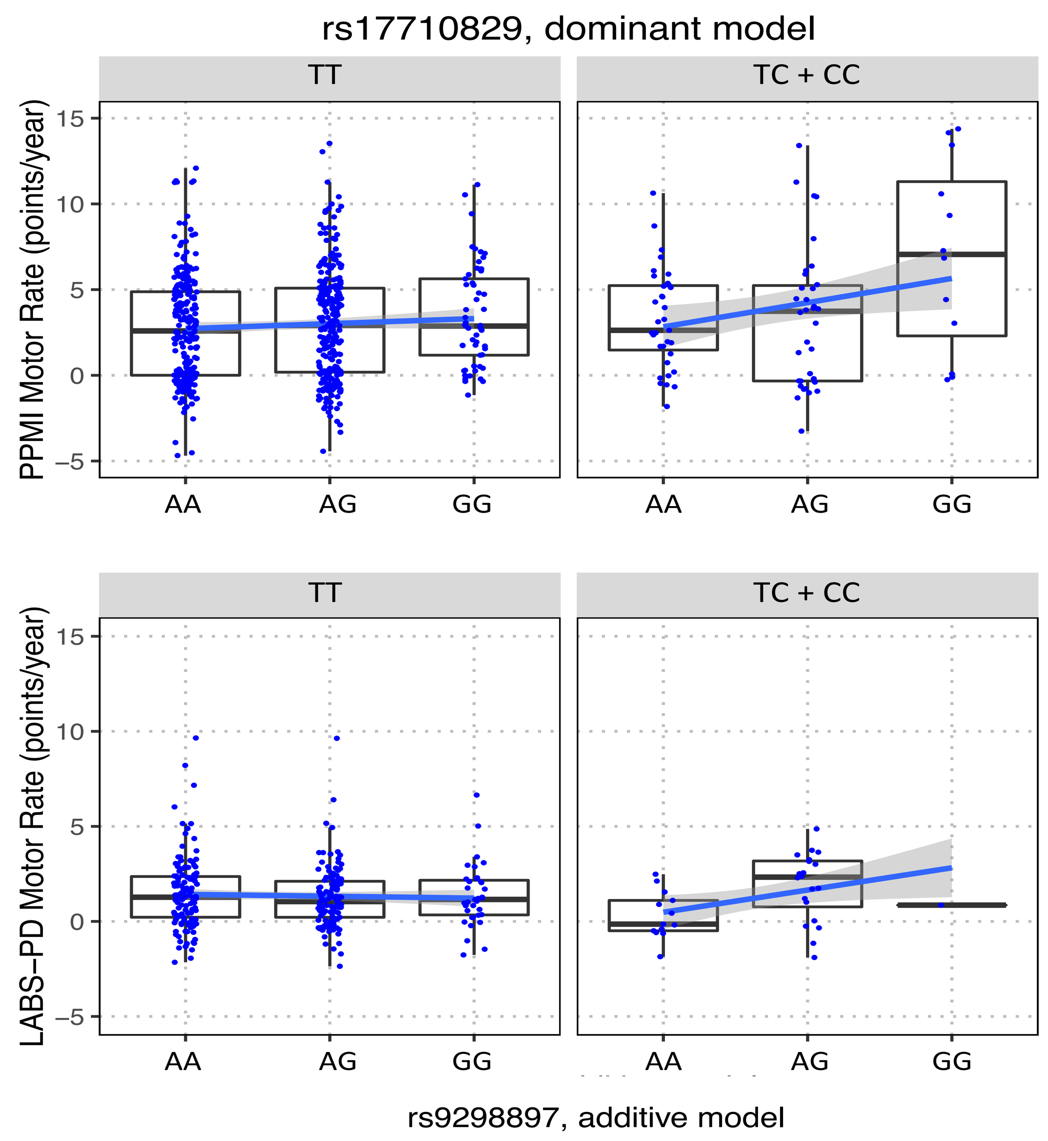

Figure 3. Cases with minor allele for both SNPs have faster decline in motor scores in both PPMI (average 2.4 **MDS-UPDRS Part II & III units/years**/year) and LABS-PD (1.2 points/per year)

Conclusions

This study highlights the utility of ensemble prediction models to capture the complex interplay of clinical, genetic, and molecular profiles in disease progression.

- REFS identified Bayesian models that combined established and novel patient factors to predict progression for both motor and cognitive deficits.
- Models allow early detection of patients most likely to have rapid disease progression (Figure 2), enabling more effective trial recruitment and clinical disease management.
- Able to identify and replicate a novel genetic interaction (Figure 3), providing potential mechanistic insight into the disease process.

Acknowledgments

This work was supported by grants from the Michael J. Fox Foundation for Parkinson's Research and National Institute of Neurological Disorders and Stroke (1P20NS092529-01)

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Avid Radiopharmaceuticals, Biogen, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, MesoScaleDiscovery, Pfizer, Piramal, Roche, Servier, UCB, and Golub Capital.

References

1. Marek K, Jennings D, Lasch S, et al. The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol. 2011
2. Ravina B, Tanner C, Dieuliis D, et al. A longitudinal program for biomarker development in Parkinson's disease: a feasibility study. Mov Disord 2009
3. Xing H, McDonagh PD, Bienkowska J, et al. Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. PLoS Comput Biol. 2011
4. Friedman N, Koller D. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. Mach Learn. 2003