Description of Additional Supplementary Files

Supplementary Data 1- Manual filtering of phecodes to arrive at 112 included phecodes.

Supplementary Data 2- Performance of preliminary and final models for 112 phecodes.

Supplementary Data 3- Robustness of final models to feature selection.

Supplementary Data 4- Comparison of final models trained using different algorithms.

Supplementary Data 5- Performance of biomarker predictors for 13 phecodes.

Supplementary Data 6- Odds ratios for phecode diagnosis per quintile increase in phecode diagnosis model score.

Supplementary Data 7- Most important features for each phecode diagnosis model.

Supplementary Data 8- Hazard ratios for all-cause mortality among all participants, only cases, and only controls.

Supplementary Data 9- Number of genes identified by P, B, and C for each phecode.

Supplementary Data 10- Highest-scoring gene-phecode pairs without a drug indication or target-disease associations.

Supplementary Data 11- Highest-scoring gene-phecode pairs with > 30% increase in score from L2G + Clinical + P to ML-GPS.

Supplementary Data 12- Highest-scoring gene phecode pairs overall.

Supplementary Data 13- Normalized enrichment scores for all hallmark gene set-phecode combinations.

Supplementary Data 14- Hallmark gene sets enriched with ML-GPS but not with L2G + Clinical + P.

Supplementary Data 15- Missingness rates of 72 laboratory and vital measurements used for phecode prediction models.

Supplementary Data 16- Differences in proportion of cases between participants with and without GP records.

Supplementary Data 17- Differences in feature values between participants with and without GP records.

Supplementary Data 18- Probability thresholds yielding maximum F1 scores for each phecode.

Supplementary Data 19- Assignment of independent genome-wide association loci to genes.