NeoGx: Machine-Recommended Rapid Genome Sequencing for Neonates

Austin A. Antoniou[1,2], Regan McGinley[3], Marina Metzler[4,5], Bimal P. Chaudhari[2,3,6,7,8]

[1]*The Office of Data Sciences, The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, USA*

[2]*The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA*

[3]*Division of Genetic and Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA.*

[4]*Division of Newborn Medicine, Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA*

[5]*Division of Newborn Medicine, Women and Infants Center, St. Louis Children's Hospital, St. Louis, MO, USA*

[6]*Division of Neonatology, Nationwide Children's Hospital, Columbus, OH, USA.*

[7]*Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio, USA*

[8]*Center for Clinical and Translational Science, The Ohio State University and Nationwide Children's Hospital, Columbus, OH, USA.*

ABSTRACT

Background

Genetic disease is common in the Level IV Neonatal Intensive Care Unit (NICU), but neonatology providers are not always able to identify the need for genetic evaluation. We trained a machine learning (ML) algorithm to predict the need for genetic testing within the first 18 months of life using health record phenotypes.

Methods

For a decade of NICU patients, we extracted Human Phenotype Ontology (HPO) terms from clinical text with Natural Language Processing tools. Considering multiple feature sets, classifier architectures, and hyperparameters, we selected a classifier and made predictions on a validation cohort of 2,241 Level IV NICU admits born 2020-2021.

Results

Our classifier had ROC AUC of 0.87 and PR AUC of 0.73 when making predictions during the first week in the Level IV NICU. We simulated testing policies under which subjects begin testing at the time of first ML prediction, estimating diagnostic odyssey length both with and without the additional benefit of pursuing rGS at this time. Just by using ML to accelerate initial genetic testing (without changing the tests ordered), the median time to first genetic test dropped from 10 days to 1 day, and the number of diagnostic odysseys resolved within 14 days of NICU admission increased by a factor of 1.8. By additionally requiring rGS at the time of positive ML prediction, the number of diagnostic odysseys resolved within 14 days was 3.8 times higher than the baseline.

Conclusions

ML predictions of genetic testing need, together with the application of the right rapid testing modality, can help providers accelerate genetics evaluation and bring about earlier and better outcomes for patients.

BACKGROUND

Genetic diseases are a leading cause of infant morbidity and mortality. As such, they are also a common cause of Neonatal Intensive Care Unit (NICU) admission, particularly level III and IV NICUs[1–4]. By far, the most common are easily recognizable, classical, aneuploidies such as trisomies 13, 18, and 21[5–7]. However, there are more than 10,000 genetic disorders which, while individually rare, are collectively common[8,9]. Each of these may manifest as a complex phenotypic presentation and may not be easily recognizable by neonatologists or even geneticists who have not been trained to recognize the neonatal manifestations of many disorders, particularly in premature infants[10]. Indeed, even though genetic disorders are relatively common in NICUs, many patients do not receive a diagnosis until after discharge[11,12]. In such a high acuity care setting and in an era where precision therapies are increasingly available[13], there is urgent need to shorten the diagnostic odyssey for neonates admitted to the NICU.

Exome Sequencing (ES) and Genome Sequencing (GS) are broad genetic tests capable of capturing the genomic variation responsible for most genetic diseases for which an etiology is known[14]. Rapid GS (rGS) has been implemented in several patient care settings with turnaround time of <7-10 days[15]. Providers and patient families alike have reported high utility, even in cases where no diagnosis was made[16]. While rGS has a higher unit cost than narrower tests, the broader, more comprehensive nature of such testing makes it a more cost-effective choice in many settings, particularly early in a NICU course[17,18].

Much of the current burden of making genetic testing decisions in the NICU – particularly consulting the Genetics service – lies with neonatology providers, who may not be ideally equipped, or confident in their ability, to make these decisions[19,20]. It is at this stage that providers and patients could benefit from machine-aided recommendations for who is likely to need genetic testing in the future. Additionally, because, in many states, the upfront cost of the genetic testing is absorbed by the NICU but the benefits of cost saving accrue to the payor, many centers engage in utilization review which either limits access to all genetic testing or restrict testing to narrower tests[21]. Because of the paucity of experts in genetics and neonatology available at many centers to lead efforts to optimize patient selection for early, broad genetic testing, we developed a Machine Learning (ML) model to predict the need for early-in-life genetic testing based on phenotypic data harvested from the Electronic Health Record (EHR).

Compared to previous efforts in this area, we study a larger cohort of neonates, investigate more complex model architectures and feature sets, and set our sights on a fundamentally different predictive target. Rather than training a model training to predict diagnoses or to replicate existing provider decisions to pursue a specific test such as microarray[22], ES, or GS[23], we set out to train the model to recognize the phenotypic presentations of patients who go on to receive a wide range of genetic tests in the hope that accurate predictions can, in future studies, be used to inform high utility, cost effective, genetic testing practices in Level IV NICUs.

METHODS

Cohort

Our study population consisted of N=33,315 patients admitted to a Nationwide Children's Hospital (NCH) Neonatal Network NICU who were born between 1 January 2010 and 31 December 2021. The NCH Neonatal Network consists of 6 Level III NICUs at delivery centers and 1 Level IV NICU in central Ohio. Neonates in NICUs in the Neonatal Networked are cared for Neonatologists from 3 practices (1 academic, 2 private) and a common pool of Neonatal Nurse Practitioners and Physicians' Assistants using a shared EHR with clinical and administrative data collected in a single Research Data Warehouse (RDW). Within the neonatal network, patients are assigned a unique identifier which allows self-self linkage of data from delivery centers with data from the Level IV NICU. Patients admitted to the Level IV NICU from outside of the Neonatal Network are not linked to data from the referring center aside from basic demographic and administrative data including birthweight and gestational age at birth. Patients who return to NCH for post-ICU care (including genetics referral/consultation or genetic testing ordered within the NCH EHR) have their data captured in the RDW with self-self linkage to the data from the NICU hospitalization.

Data Acquisition & Study Definitions

For each subject, birthweight, gestational age at birth (GA), all available lab tests, genetics consults and referrals, and International Classification of Disease (ICD) 9 and 10 codes for the length of their entire health record were obtained from the NCH RDW. Clinical note text, when available, from the NICU was processed by ClinPhen[24] to produce a set of Human Phenotype Ontology (HPO)[25] terms to represent the phenotypic profile of each patient. While metadata such as age when a specific phenotype or ICD code was first noted in the medical record was available, each subject was assigned a single phenotypic profile summarizing their entire NICU course. Subjects readmitted to the NICU only had their initial hospitalization characterized in this way.

Lab tests were marked as *genetic* or *non-genetic*, with the "genetic" class including ES/GS, single gene or panel next-generation sequencing (NGS) tests, microarrays, karyotypes, Fluorescent In Situ Hybridization, methylation, repeat, and uniparental disomy testing.

We identified patients in our study cohort with at least one of the following in the NCH medical record, appearing before 18 months of age, for further review: (1) an order for a genetic test, (2) a consult with a medical geneticist, or (3) an ICD code signifying a disease with known genetic etiology, including classical autosomal aneuploidies. The charts of patients who had (2) or (3) but no genetic testing in our medical record system, were reviewed to assess whether they received genetic testing elsewhere. Subjects with a genetic test were given an outcome label of *genetic*. All other subjects were labelled *non-genetic*. The 18-month threshold for initial suspicion of genetic diagnosis was chosen under the rationale that their eventual genetic evaluation might have been motivated by common developmental and

behavioral phenotypes, like autism, rather than phenotypes related to the underlying reasons for their NICU hospitalization.

Within the genetic group, we made the additional determination of whether each subject had a confirmed diagnosis as opposed to simply a history of non-diagnostic testing. Specifically, diagnoses of trisomy 13, 18, or 21 were found by searching History and Physical Examination (H&P) notes to identify patients with positive prenatal screens or diagnostic tests, which are frequently coded in the mother's health record and not the newborn's. Regular expressions were used to match phrases indicating (1) diagnosis or positive prenatal screen and (2) trisomy 13, 18, or 21, or any of the corresponding syndrome names for these diseases. Any patient with a sentence in an H&P note matching (1) and (2), subject to subsequent chart review, was also labeled as having a common autosomal aneuploidy.

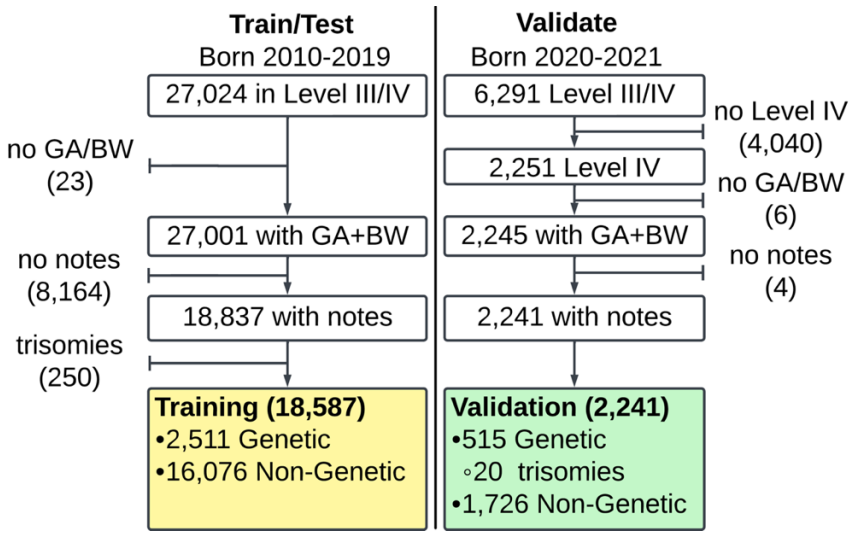

Figure 1: Modified CONSORT/Study Flow Diagrams for Training and Validation Cohorts. *The full data set was split into subjects born between 2010 and 2019 (for model selection and training) and subjects born between 2020 and 2021 (for model validation). Each step in the diagram shows how many subjects met (and how many failed to meet) the inclusion criteria.*

Cohort Description		Training	Validation
		18587	2241
Sex, n (%)	Male	10358 (55.7)	1252 (55.9)
	Female	8229 (44.3)	989 (44.1)
Race, n (%)	White	12058 (64.9)	1545 (68.9)
	Black or African American	3419 (18.4)	367 (16.4)
	Unknown	1230 (6.6)	107 (4.8)
	Multiple race	1273 (6.8)	163 (7.3)
	Asian	533 (2.9)	54 (2.4)
	Other	39 (0.2)	
	American Indian or Alaska Native	19 (0.1)	5 (0.2)
	Native Hawaiian or Other Pacific Islander	16 (0.1)	
Ethnicity, n (%)	Not Hispanic or Latino	17148 (92.3)	2101 (93.8)
	Unknown	937 (5.0)	74 (3.3)
	Hispanic or Latino	502 (2.7)	66 (2.9)
GA (wks), mean (SD)		35.1 (4.5)	35.3 (4.7)
BW (kgs), mean (SD)		2.5 (1.0)	2.5 (1.1)
Outcome, n (%)	Genetic Test or Dx (Non-trisomy)	2511 (13.5)	495 (22.1)
	Negative	16076 (86.5)	1726 (77.0)
	Trisomy		20 (0.9)
Known Dx, n (%)		685 (3.7)	156 (7.0)
Known GS, n (%)		245 (1.3)	105 (4.7)

Table 1: Training and Validation Cohort Summary. *For categorical variables, the counts and percentage of each within the training and validation cohorts, respectively, are shown. For continuous variables, the mean and standard deviation within the training and validation cohorts are shown. (wks: weeks; kgs: kilograms; SD: standard deviation; GA: gestational age; BW: birth weight; Dx: diagnosis; GS: genomic sequencing).*

Model Development

Training-validation split

Training and validation data were separated from each other by splitting the cohort, based on birthdate, at the date January 1, 2020 (Figure 1). The training set consisted of 18,587 NCH NICU patients born before 2020 for whom clinical text, birth weight (BW), and gestational age (GA) data were available, after excluding patients with known common autosomal aneuploidy. For validation, 2,241 patients born on or after January 1, 2020, and for whom all data were available, were set aside. The *genetic* and *non-genetic* labels corresponded to the positive and negative classes for ML classification, though patients with a known common autosomal aneuploidy were excluded from model training. Patients diagnosed with

common autosomal aneuploidies were included in the positive class of the validation cohort; however, as a sensitivity analysis, we also calculated classification performance metrics excluding predictions made for these patients.

Model features & Feature Engineering

BW was converted to Fenton 2013 sex and GA-adjusted Z-score using the PediTools bulk calculator[26]. GA was divided into bins bounded by 28, 32, 35, 37 completed weeks of gestation. Additional features to represent phenotypes were calculated from each patient's set of EHR text-derived HPO terms (Figure 2). The HPO consists of more than 16,000 terms, making representation of each term infeasible for machine learning given our sample size. We instead chose a representative set of terms to act as a feature set, selecting terms for the density of information contained in their descendant terms in the HPO graph. We explored, but ultimately did not use, representations based on depth below the root node of the HPO directed acyclic graph, as well as encoding of feature values as present or absent. Here we describe the approach which was ultimately selected (See supplement for details on feature sets and models considered but not selected).

We borrowed the formulation of Information Content (IC) from Phrank, derived from HPO's phenotype-gene associations, to act as a value function for sets of HPO terms[27]. The Information Potential Ratio (IPR) of a term x, meant to encode the density of information in the set of terms logically entailing x relative to the specificity of x itself, was calculated as

$$IPR(x) = \frac{IC\big(\text{descendants}(x)\big)}{IC(x)}$$

We chose as our representative set the collection of terms with $IPR(x) \geq 10$, accounting for hierarchical redundancy by removing any terms in F with descendants also belonging to F. This left a set R of 77 representative terms.

Letting X_{subj} be the set of HPO terms in a subject's EHR text and letting x in R be one of the representative feature terms, the subject's feature value for the term x was calculated the total IC of all terms in X_{subj} which logically entail or are logically entailed by x; that is,

$$IC_x^{\text{bidirectional}}(\text{subject}) = IC(X_{\text{subj}} \cap \text{descendants}(x) \cap \text{ancestors}(x))$$

With 77 features encoded as above along with the male sex binary indicator, the one-hot encoded GA bins, and the Fenton 2013 BW Z-score, the model was provided with 83 input features.

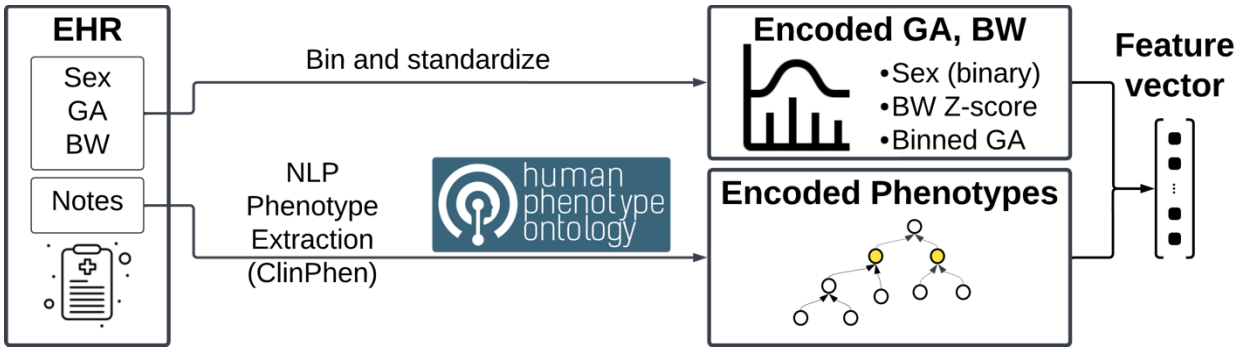

Figure 2. Creation of ML features from raw input data. *Each raw data element undergoes preprocessing to make it usable as ML input. Biological sex is a categorical field which is encoded for the model as a one-hot encoded indicator of being male or not. Gestational Age at birth (GA), measured in weeks, is categorized as belonging to at most one of the intervals [0,28), [28,32), [32,35), [35, 37), and indicator variables are input to the model accordingly. Birth weight (BW) is transformed to a sex and GA-adjusted Fenton 2013 Z-score. Clinical notes are processed by ClinPhen, which extracts HPO terms. The set of all extracted HPO terms is then rolled up by logical entailment to any selected terms in the set of 77 representative feature terms, with value equal to IC of the set of terms associated to each representative feature term.*

Model selection

A grid search was carried out in scikit-learn using 10-fold cross validation on the training set. The grid search considered feature sets, feature encodings, four classifier architectures – Naïve Bayes, Logistic Regression, Random Forest, and XGBoost – and classifier specific hyperparameters. Models were evaluated on the fold mean of their Precision Recall (PR) area under curve (AUC) cross-validation score. To balance model performance on the training set with overall complexity of the model and feature set complexity, we a priori determined to select the most parsimonious model with training PR AUC scores within 0.01 of the best-performing model. The final selected classifier was retrained on the full training data set, using all phenotype data available through the end of the NICU stay. The classifier was then isotonically calibrated and the Brier loss was calculated before and after calibration.

Measuring classification performance

We evaluated the model's ability to predict genetic testing need by simulating the accrual of phenotypic information over the first 7 days of NICU stay for all subjects in the test set and accumulating the predictions made by the model on each day during this period. For each day in the level IV NICU, we truncated the feature matrix to include only information available for subjects remaining in the Level IV NICU at that time. Each subject's final prediction probability was equal to the maximum of the classifier's prediction probabilities over the 7-day simulation. The Receiver Operator Characteristic (ROC) AUC and the PR AUC scores were calculated for the classifier's predictions. To account for the effect of the classical aneuploidies on positive predictions, we also report AUC scores for the validation cohort after excluding these patients. Additionally, to mitigate the effect of skew on minimum achievable PR AUC, we report a normalized PR AUC score[28].

Each classifier considered returns a calibrated confidence score in the interval [0,1] which is the predicted probability that the given patient will receive genetic testing or obtain a genetic diagnosis within the first 18 months of life. A particular recommendation model can be derived from the classifier by fixing a probability threshold τ; that is, the model recommends a test when the classifier probability is $\geq \tau$. The ROC AUC and PR AUC scores mentioned above account for this, measuring classifier performance across the full range of decision thresholds. For the evaluation of bias and estimates of benefit, where binary predictions are required, we fixed a probability threshold which yielded maximum F1 score on the validation set (other thresholds are considered in the supplement).

Estimation of benefit and cost

We measured the potential impact of model recommendations on diagnostic odyssey length by simulating different clinician responses to the recommendations. The *diagnostic odyssey length* was defined as the number of days from Level IV NICU admission to last known genetic test result or molecular genetic diagnosis, censored at 18 months. Subjects who had testing but never received a diagnosis or had ES/GS were considered not to have completed their diagnostic odyssey. Subjects who initiated testing before 18 months but whose odysseys continued beyond 18 months were censored at 18 months.

The diagnostic odyssey can be thought of as including two segments: the *initiation time*: the time from Level IV NICU admission to the first test ordered*;* and the *testing duration*: the time from first test order to final test result, which may span months or years. The first segment can be shortened by encouraging early testing, and the second can be shortened by encouraging rapid testing such as rGS.

We simulated 5 policies for applying ML recommendations and rGS to explore the relationship between these aspects of shortening diagnostic odysseys. For simulations including rGS, we approximated rGS turnaround time from sample collection to final result as an exponential random variable with a rate parameter of 5 days, based on turnaround times observed from previous studies of rGS implementation[29].

(1) *Actual odyssey* served as a baseline, modeled using the actual diagnostic odyssey length measured from patient data.

(2) *ML-initiated testing* only changes the initiation time, reducing the time to first test for subjects who received a positive ML prediction while not altering the testing duration (test initiation time remains the same for subjects who received a negative predictions).

(3) *ML-initiated rGS* changes the initiation time and replaces all testing with rGS for all subjects with a positive ML recommendation.

(4) *Only rGS, no ML* was modeled by leaving the test initiation time unchanged but replacing all testing with rGS.

(5) *Only rGS + ML* supposed that every subject received rGS as their first and only test, but patients with positive ML predictions started testing on the day of their positive prediction.

For policies (2) and (3), the odysseys of subjects with negative ML predictions were left unchanged from the actual diagnostic odyssey, so only subjects who received positive predictions could receive benefit. We defined "completion of diagnostic odyssey" as a subject receiving either a diagnosis or genomic sequencing test.

We assessed the isolated effect of ML recommendations on shortening time to test initiation by measuring median and IQR test initiation times with and without ML recommendations. To test the combined effect and pairwise differences between the policies (1)-(5) outlined above, we used Kaplan-Meier estimators right-censored at 18 months, log-rank tests for significance, as well as pairwise McNemar tests for the significance of differences in the number of subjects completing diagnostic odyssey within 14 days of Level IV NICU admission under each policy.

Evaluation of model bias

Bias relative to patient characteristics: sex, race, ethnicity, gestational age

Bias in this problem may stem from the underlying representation of demographic groups in our data, genetics utilization practices to date and resulting consequences on labelling, or the predictive behavior of the model itself[30]. We considered the impact of these factors across subpopulations defined by the characteristics: sex, race, ethnicity, and GA[12].

To evaluate biases in the underlying data, we stratified the validation cohort by each population characteristic, making pairwise comparisons of the 95% confidence intervals for each subpopulation. We deemed a subpopulation difference to be significant if the intervals were entirely disjoint; otherwise any differences were not considered significant.

Biases related to the model's predictive behavior were assessed by a similar comparison between each subpopulation's confidence intervals for the proportion of positive ML predictions, precision, and recall[30].

We also assessed the differences in test initiation times for the genetic subgroups of each subpopulation by comparing confidence intervals for the actual time at which testing was initiated (Policy 1) and for the ML-modified testing initiation time (Policies 2, 3, 5).

Bias relative to patient outcomes: diagnosis and genomic sequencing utilization

As previous work in this area has focused on predicting whether a child has a genetic diagnosis[22,31] or received GS[23], we also sought to characterize the model's performance with respect to these targets. We followed the same line of analysis for differences observed between subgroups of subjects who experienced different outcomes. Specifically, we split the genetic group in the validation cohort with respect to whether a subject received genomic sequencing and whether they ever received a genetic diagnosis. To examine interaction effects, we simultaneously split by both these outcomes. Because these are all subpopulations of the genetic group, only recall and test initiation times could be sensibly compared between outcome subgroups.

Feature importance

SHapley Additive exPlanations (SHAP) values were calculated to measure the importance of features in making positive and negative predictions[32]. The top features were determined to be those with highest mean absolute SHAP values taken over all subjects in the validation cohort. Furthermore, at a fixed classification threshold, SHAP values were calculated for the subsets of the validation cohort corresponding to false positive and false negative predictions and top features were recalculated for these subsets to probe the relationships between feature importance and specific failure modes.

RESULTS

Selected classifier

The model from the parameter grid search with the best 10-fold averaged PR AUC score (0.703) was an XGBoost Classifier with a set of 441 IC-encoded HPO features, corresponding to an IPR threshold of 10. However, an XGBoost with 83 total features (77 IC-encoded HPO features at an IPR threshold of 100) achieved a fold-averaged PR AUC score of 0.70 and ROC AUC or 0.89. This was the smallest feature set which yielded a training PR AUC score within 0.01 of the best-scoring classifier, so the corresponding model was chosen. This model was retrained on the complete NICU stays of all subjects in the training set and then isotonically calibrated, bringing its Brier loss score from 0.11 to 0.10.

Classification performance

As the predictions included information ranging from 1 to 7 days in the Level IV NICU, performance of the model slightly improved, with test set ROC AUC scores rising from 0.86 to 0.87, and PR AUC score rising from 0.72 to 0.73 (Figure 3). Excluding the 20 patients with classical aneuploidies did not materially affect these results (see supplement). The classifier's skew-normalized PR AUC was 0.69.

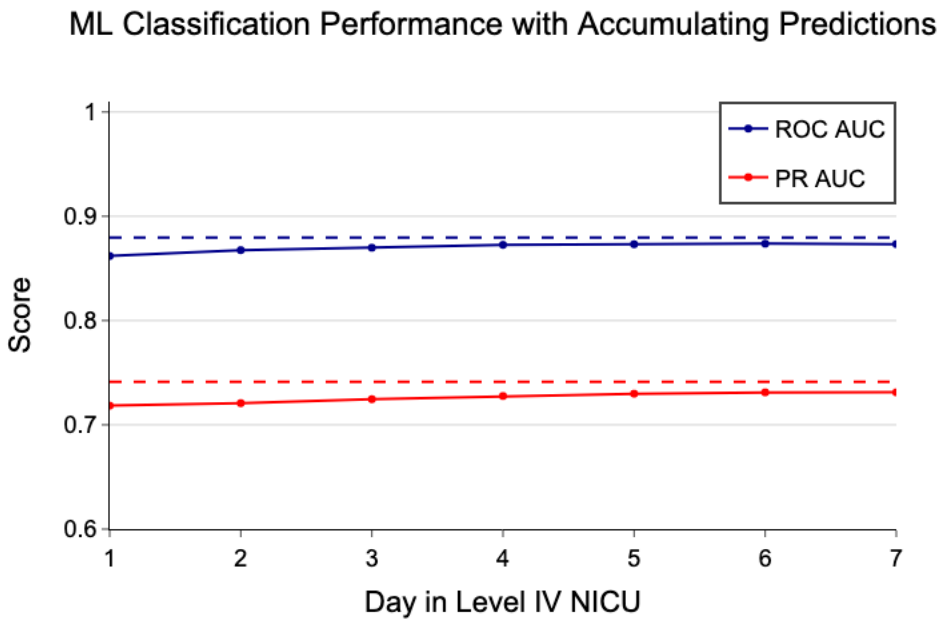

Figure 3. Prediction performance at each timestep. *For the first 7 days in the Level IV NICU, predictions were made at each day for each subject based on the maximum probability score calculated by the classifier through the current day. The blue and red curves represent ROC and PR AUC scores, respectively, at the given prediction timepoint, where each subject's prediction probability on a given day was equal to the maximum of prediction probabilities on that day and all previous level IV NICU days. The dashed lines indicate the score using predictions made with phenotypic information from the full NICU stay.*

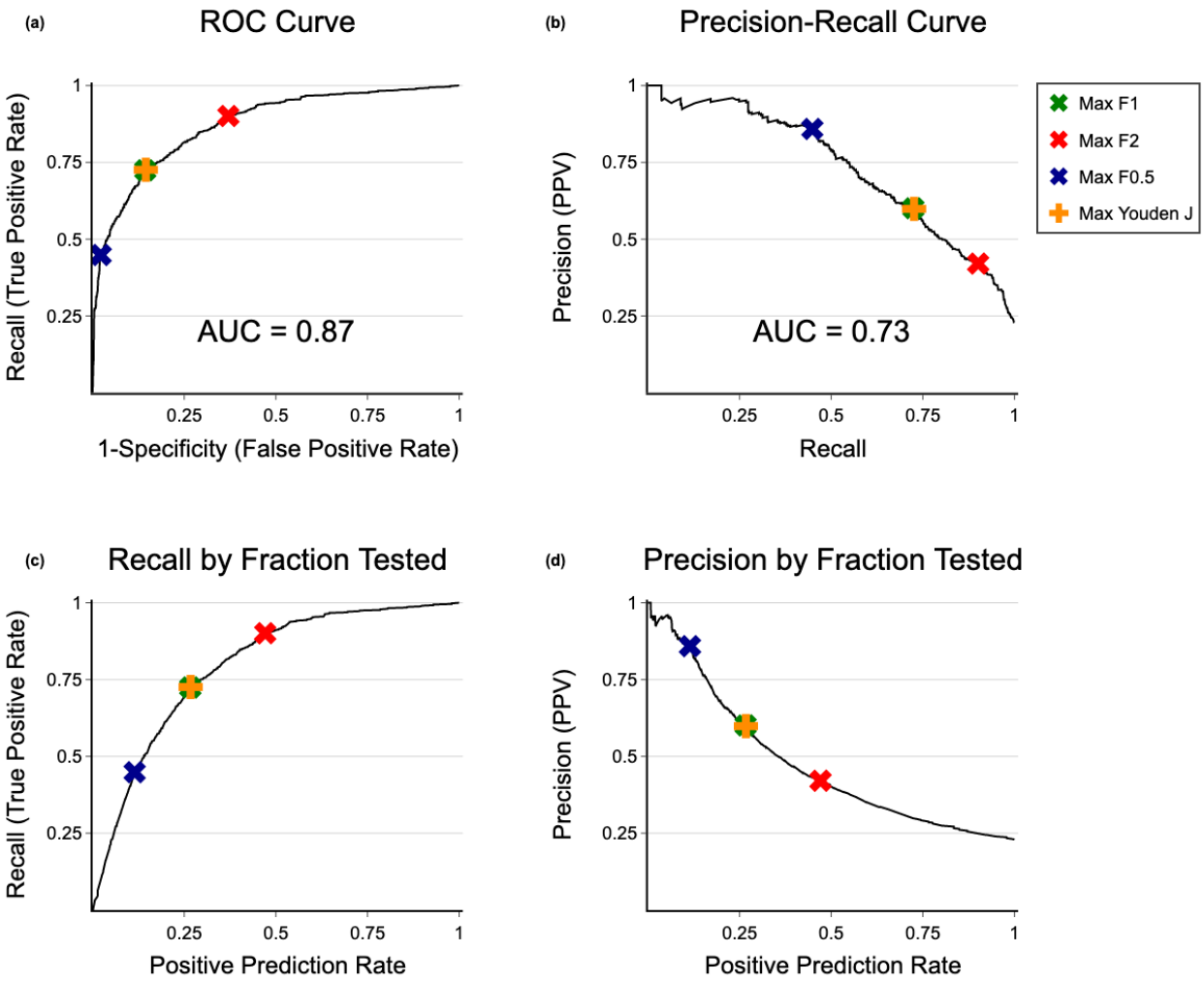

Figure 4. Classification curves for 7-day accumulating predictions*. (a) ROC curve; (b) PR Curve; (c) Recall versus positive prediction rate; (d) Precision vs overall positive prediction rate. Markers indicate decision boundaries maximizing validation cohort F1, F2, F0.5, and Youden J scores.*

For the remainder of the analyses where binary predictions are required, we report those made by the F1 model which recommends testing when the predicted probability is at least 0.21. Other potential thresholds for binary classification and their implications for true and false positive rates are shown in Figure 4.

Simulated diagnostic odysseys and estimated benefit

Reduction of Time to Initial Testing with ML Recommendations

We compared the actual time at which initial genetic testing was ordered to the hypothetical time at which the ML model would have recommended testing (Figure 5). Applying ML recommendations dropped the median test initiation time from 10 days to 1, and the 75th %-ile initiation time from 54 to 2 days ($p<0.01$ for both differences).

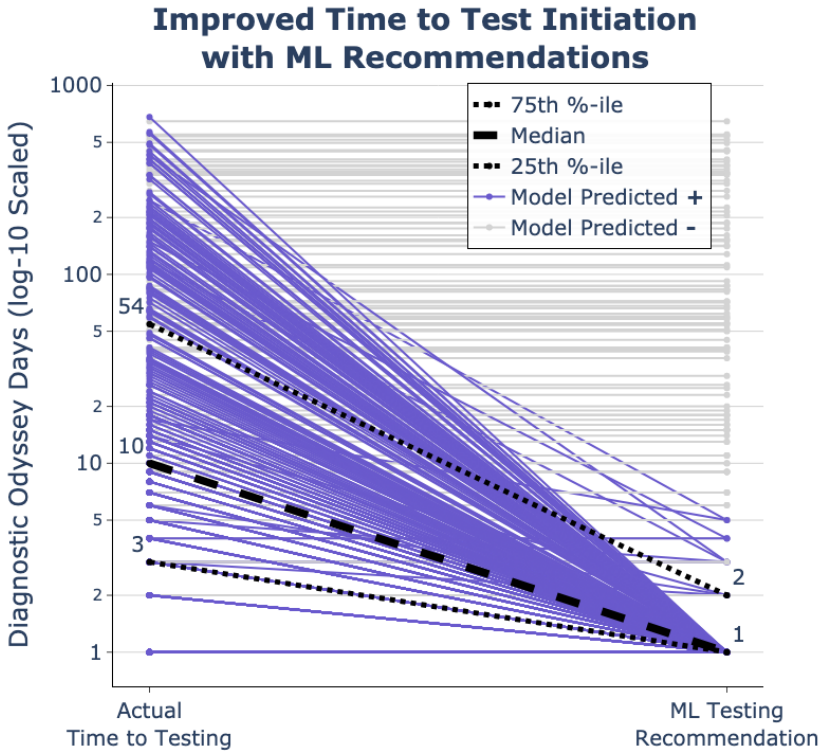

Figure 5. Parallel plot comparing test initiation times with and without ML. *The points on the left show the day of Level IV NICU hospitalization on which each subject's first test was ordered. The points on the right show the hypothetical time at which each subject would initiate testing under the recommendations of the ML model. The line connecting a subject's actual time to their hypothetical time is blue if the ML model made a positive prediction; otherwise the line is grey.*

Simulated time to diagnostic odyssey conclusion

Compared to actual diagnostic odysseys (Policy 1 – Actual Odyssey, Methods), which were concluded within 14 days in only 15.2% of actual cases, initiating testing based on ML recommendations (but not changing any of the test selection; Policy 2 – ML-Initiated Testing, Methods) increased this to 28.3% ($p<10^{-30}$). Enacting Policy 3 (ML-initiated rGS, Methods) led to completion of diagnostic odysseys withing 14 days in 72.4% of cases ($p<10^{-22}$ vs. Policy 2). Under Policy 4 (Only rGS, no ML), only 49.5% of diagnostic odysseys are completed within 14 days ($p<10^{-24}$ vs Policy 3). Finally, an "optimal" policy where the only genetic test ordered is rGS and testing initiation is informed by ML predictions (Policy 5, Methods) increases completion of diagnostic odysseys by 14 days to 79.0% ($p<2.5\times10^{-4}$ vs Policy 4). A

comparison of the Kaplan-Meier curves for the strictly comparable Policies 1, 2, 3 and 5 by log rank test demonstrates that Policy 2 is superior to Policy 1 ($p<5\times10^{-8}$) and Policies 3 and 5 are superior to Policy 2 ($p<10^{-6}$) but they are not statistically different from each other (Figure 6).

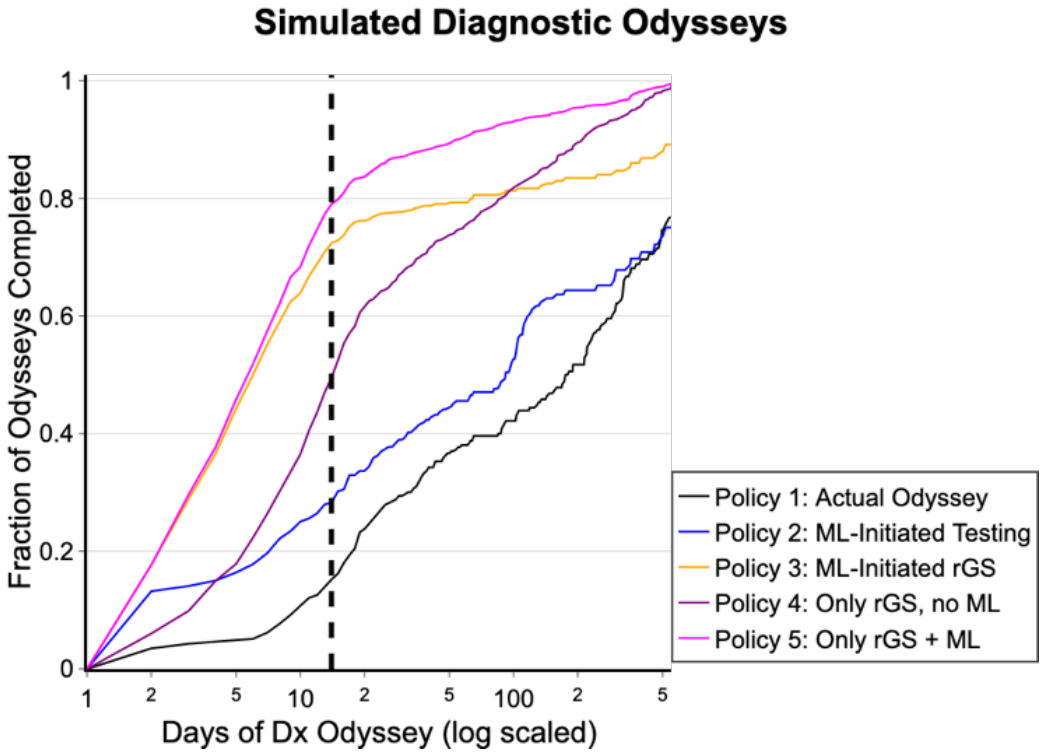

Figure 6. Time to diagnostic odyssey completion under simulated recommendation policies. *Under the three policies – no model, ML-initiated testing, and ML-initiated rGS – we measure the time taken for all positively labelled subject in the test set to complete their diagnostic odyssey. The horizontal axis measures the number of days after Level IV NICU admission, log scaled. The vertical axis measures the fraction of all positively labelled subjects in the test set – those with known genetic testing or diagnoses – who have received a genetic diagnosis or GS test by the specified timepoint. The different policies include: Actual odyssey (black); ML-initiation (blue) doesn't change the tests ordered, just the time at which testing starts; ML-initiated rGS (orange) assumes rGS is ordered at the time a positive prediction is made; only rGS with no ML (purple) replaces all initial test orders with rGS but does not change the time at which the test was ordered; only rGS with ML (pink) replaces all initial orders with rGS, and accelerates the initial order time in the case of positive predictions.*

Impact of recommendations on test utilization

Considering Policy 3 as the best performing policy, we consider testing utilization changes. Under Policy 3, 621 patients, or 27.7% of the validation cohort are recommended for rGS. However, these recommended rGS would replace tests for 373 patients who did undergo genetic evaluation (precision=0.60), including 74 karyotypes, 184 microarrays, and 187 single gene or gene panel next generation sequencing tests. Additionally, 93 of these individuals would have gone on to receive ES/GS without the model's recommendation. Notably, the model recalled 19/20 patients with a known classical aneuploidy. Of these, 18 were already diagnosed by day 1 of patient's Level IV NICU stay, and thus received no benefit from the rGS recommendation.

The 621 recommendations for testing under Policy 3 include 248 potentially erroneous test recommendations (False Positives) for patients who did not otherwise have any involvement with genetics as prospectively defined for this study (Policy 1). However, a manual chart review of these subjects in April and May 2024 revealed 6 had begun diagnostic odysseys after 18 months of age, 10 had a confirmed molecular diagnosis or testing apparent from chart narratives, and 62 were judged by an expert in inpatient genetic consultation (BPC) to likely meet current standards for genetic testing in the study NICU. The most commonly identified recurrent cause of false positives was prematurity (N=20).

Classification bias across patient characteristics

No differences were found between the male and female subpopulations for any of the quantities measured: “genetic” prevalence, positive prediction rate, precision, and recall.

A lower positive prediction rate was noted in the unknown race subpopulation than in the Asian, Black or African American, or White subpopulations, as observed by non-overlapping confidence intervals. Otherwise, there were no significant pairwise differences between racial groups for genetic prevalence, precision, or recall.

The unknown ethnicity subpopulation had a lower rate of positive prediction than the Non-Hispanic or Latino subpopulation, but no other significant differences were observed between ethnic subpopulations for any of the other measurements.

The fraction of “genetic” subjects and the positive prediction rate were lower in patients born < 32 weeks GA than in subjects born ≥ 35 weeks. Additionally, the rate of positive predictions was lower for subjects born < 28 weeks than for all subjects born ≥ 32 weeks. No significant differences in precision were observed between GA groups, but model recall was better for subjects born ≥ 35 weeks than for subjects born < 28 weeks.

Though we have mentioned the significant pairwise differences above, all pairwise comparisons can be made from Figure 7. We additionally examined subpopulation differences in actual time to test initiation (Policy 1) and ML-initiated testing (Policy 2, 3, 5). The only significant difference between subpopulations was that the unknown ethnicity subpopulation had shorter time to testing under all policies than the non-Hispanic or Latino subpopulation. Full details of all subpopulations’ test initiation times can be found in the Supplement.

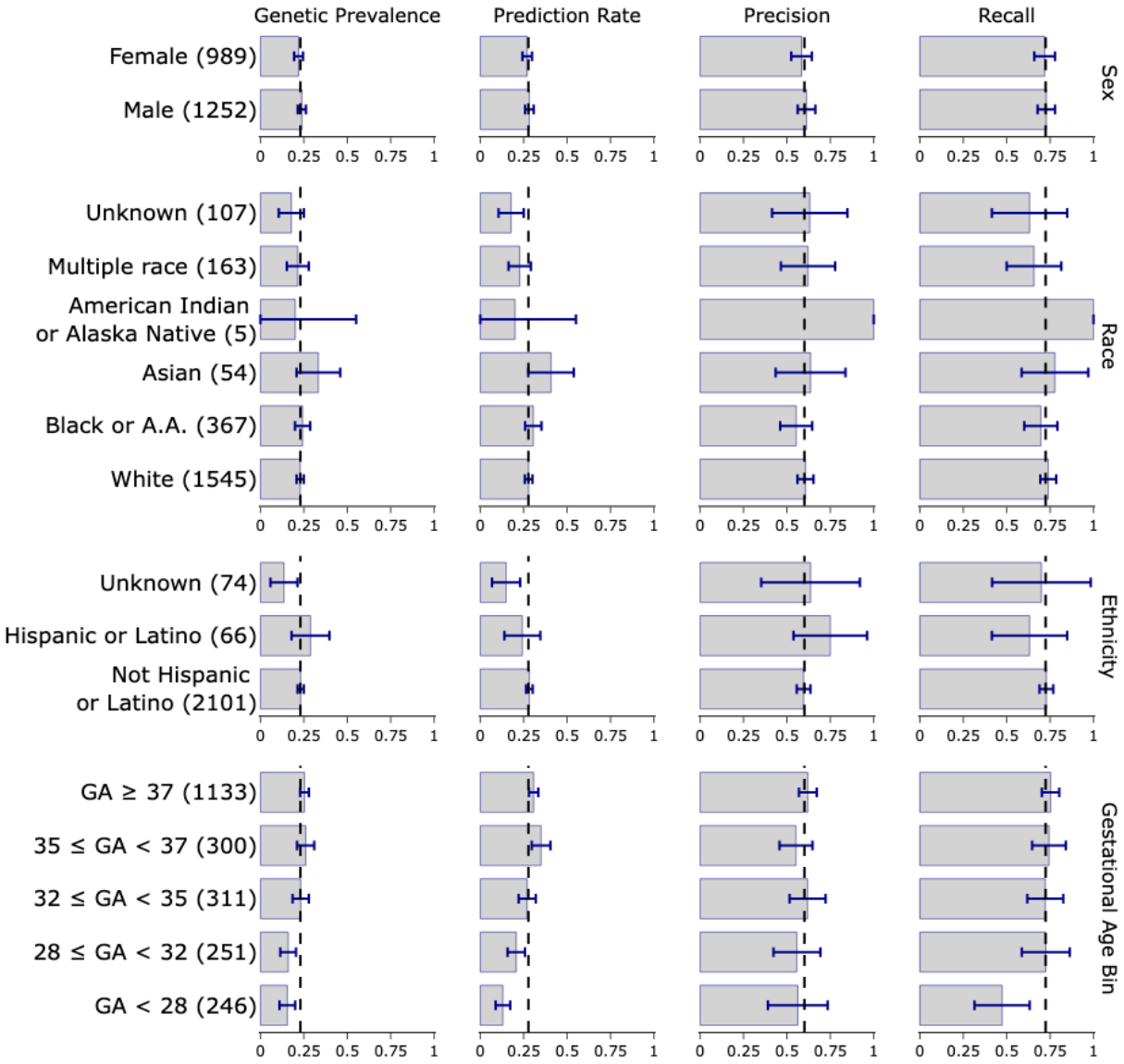

Figure 7. Bias by patient characteristics. *Shown above are 16 subplots; rows correspond to population characteristic: sex, race, ethnicity, gestational age. Columns correspond to quantity measured: genetic fraction of subpopulation, rate of positive ML prediction, precision, and recall. Each subplot (e.g. the top rightmost set of bars) shows the given metric (e.g. recall) for each subpopulation corresponding to the given patient characteristic (e.g. sex). Error bars indicate a 95% confidence interval for the given metric, and the black dashed lines show the value of each metric for the full validation cohort.*

Bias with respect to genetic outcomes

The model recalled 123/156 (78.8%) of subjects who would have eventually received a genetic diagnosis, versus 250/359 (69.6%) of subjects with genetic involvement but no known diagnosis. ML recalled 93/104 (89.4%) of subjects who had ES/GS versus 280/411 (68.1%) of subjects who did not.

We also considered recall and test initiation time while simultaneously splitting by both of these outcomes, for a total of 4 subgroups of the “genetic” group. The model recommended 191/293 (65.2%) of subjects with no known diagnosis or GS, the lowest recall among the four outcome groups. For patients

with a diagnosis but no GS, 89/118 (75.4%) were recommended by the model. The remainder of patients, who had GS testing, were the most easily recalled by the model; 34/38 (89.5%) of those with known diagnosis and 59/66 (89.4%) of those with no known diagnosis. The differences between “No diagnosis and no GS” and the groups who had GS – with or without receiving a diagnosis – were found to be statistically significant by non-overlapping 95% confidence intervals, but the remaining differences were not (see Supplement).

When considering the actual test initiation time (Policy 1, Methods), we observed that undiagnosed patients who received ES/GS had significantly longer time to test initiation than diagnosed and undiagnosed patients who had not received GS. However, under ML-initiated testing (Policy 2, 3, 5, Methods), none of the subgroups differed significantly in time to initial test order (Figure 8)

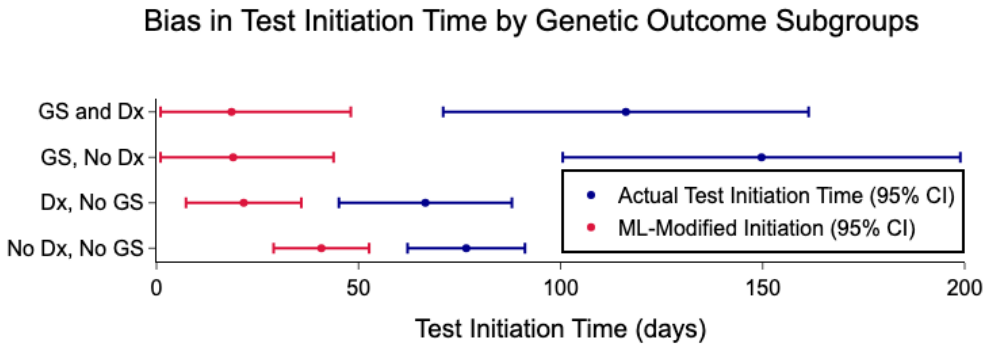

Figure 8. Test initiation time confidence intervals for genetic outcome subgroups. *For each outcome group determined by whether a subject received a diagnosis and whether they received exome or genome sequencing, 95% confidence intervals are shown in blue for their actual time from first test order (Policy 1, Methods) and in red for the time of first test order with ML-initiated testing (Policy 2, 3, 5).*

Feature importance

The features most important to informing model predictions overall, as measured by mean absolute SHAP values, were Abnormal heart morphology, abnormal facial skeleton morphology, GA < 28 weeks, abnormal pinna morphology, and abnormal nasal morphology (Figure 9). The features most important in false positive predictions included abnormal heart morphology, abnormal facial skeleton morphology, GA < 28 weeks, neurodevelopmental abnormalities, and BW z-score. False negative predictions were most informed by abnormal heart, nasal, pinna, and facial skeleton morphology, as well as GA < 28 weeks (see Supplement for SHAPs in FP and FN).

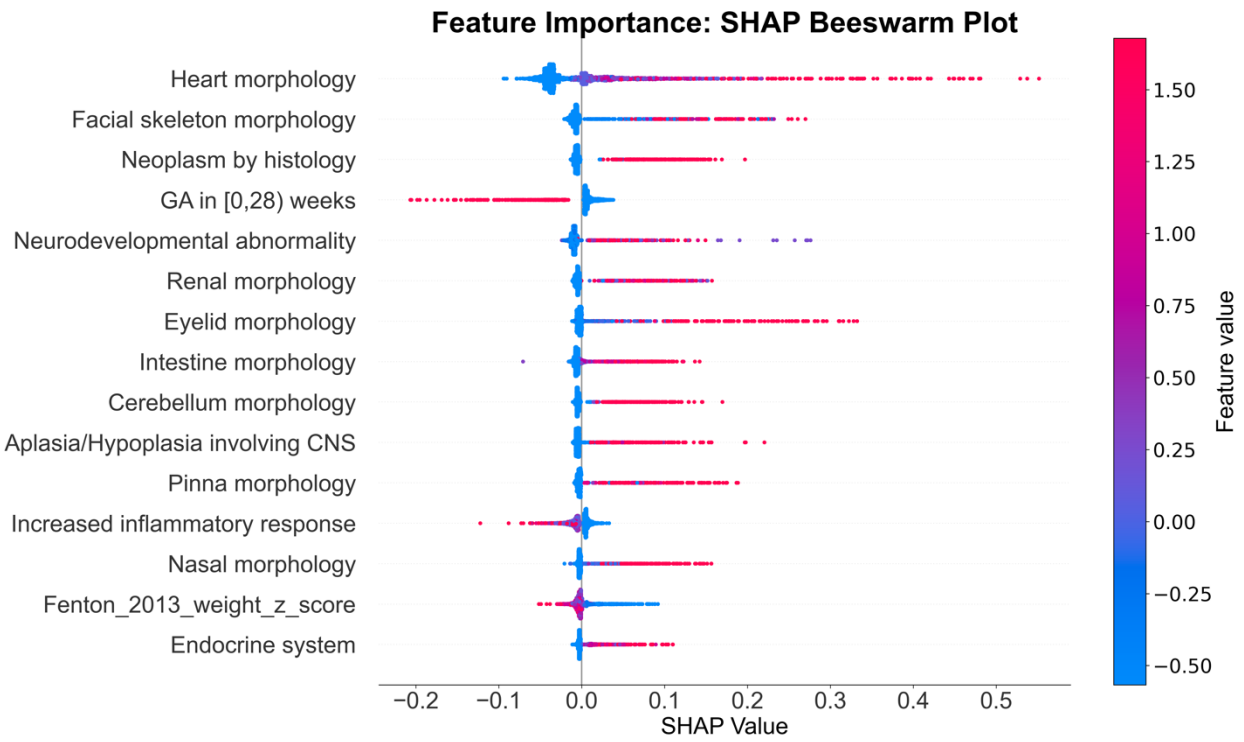

Figure 9. *SHAP Beeswarm plot. A visualization from the SHAP package of feature importances. The features of the classifier with highest mean absolute Shapley values are shown on the vertical axis, and the horizontal axis shows the SHAP value, indicating the degree to which the feature influences positive predictions. Individual points correspond to individual subjects in the test data, and the color of the point indicates whether the input value for the given feature was high or low. Feature names other than "GA in [0,28) weeks" and "Fenton_2013_weight_z_score" correspond to Human Phenotype Ontology concept labels, with "Abnormal Heart Morphology" abbreviated to "Heart Morphology", and similarly for other "Abnormality of …" or "Abnormal …" features. (CNS: central nervous system).*

DISCUSSION

Classifier Strengths and Weaknesses

A key strength of the classifier described in this work includes a novel target for supervised machine learning (future utilization of genetic testing generally as opposed to use of rGS or microarray) which, in combination with our center's robust data warehouse, allowed us to train and test our classifier on a sample an order of magnitude larger than those previously described. Importantly, because we trained on a utilization outcome, genetic testing before 18 months of age occurred at similar rates across race, ethnicity and sex. Prevalence of the target was lower amongst preterm infants with consequent lower recall. It is unclear if this difference by gestational age reflects true differences in "need" for genetic testing or bias against premature infants, though similar model precision regardless of gestational age suggests this is a real epidemiological effect rather than bias.

While no significant differences were found in the prevalence of positive outcome or in any of the model-specific scores measured (prediction rate, recall, and precision), this does not conclusively say that our model is free of bias, as some racial and ethnic groups are sparsely represented in NCH NICUs. In particular, the "American Indian or Alaskan Native" and "Asian" racial groups contained fewer than 100 people, as did the "Hispanic or Latino" and "Unknown" ethnic groups. Sample sizes this small make it difficult to extrapolate the model's predictive performance to populations consisting of a larger number of individuals from these groups.

Patients who had GS, positive or negative, were highly likely to be predicted positive by the model. This is desirable, as some of these patients were among those with the longest diagnostic odysseys, and reaching that level of testing may attest to a level of medical complexity that makes testing these patients all the more important. This also suggests a tangible phenotypic signal present in the population of patients who are likely to need sequencing eventually in life, even in phenotypes present long before GS was considered.

An additional strength of the model generated is the face validity of most important features driving recommendations. On the one hand, this is not surprising given the features were engineered largely from language used by clinicians in daily practice as well as core measure of growth and maturation used by all neonatologists and medical geneticists in their medical decision making. The most informative features, per SHAP values, correspond to HPO terms for abnormalities of organ systems which could reasonably be understood to represent congenital anomalies, though a small number of spurious features are present and worthy of discussion.

Notably, "Neoplasm by histology" appears as a high-importance feature, yet cancers and tests for diagnosing cancers are largely ignored within this study (and exceedingly rare in neonates). Upon inspection of the original clinical text from which HPO terms were extracted by ClinPhen, it seems that an overwhelming majority of contributions to this feature arose from cases in which the specialty "oncology" was mentioned in the note text and ClinPhen matched this as a synonym of the term HP:0002664

"Neoplasm". These occurrences of the word "oncology" were often present as part of a provider's signature or other oblique references to provider rather than direct mentions of cancer phenotypes in the patient. However, because Hematology-Oncology is infrequently consulted in the NICU and because such consultation is often related to concerns for genetic bone marrow failure syndromes, the signal remains informative.

Another seemingly spurious feature with SHAP values indicative of importance for negative predictions is the presence of phenotypes related to HP:0000479 "Abnormal retinal morphology". On inspection of the original set of ClinPhen-tagged HPO terms, a large fraction of the tagged terms consisted of HP:0500049 "retinopathy of prematurity", indicating that this feature serves as a proxy for prematurity. Incidentally, gestational age under 28 weeks, the feature encoding extreme prematurity directly from the structured health record, also appears as a feature with high importance for negative predictions. From this we surmise that the model has correctly learned a relationship between isolated prematurity and a lower predisposition for non-genetic underlying cause for NICU admission.

A key weakness of our classifier to consider is that we recapitulate historical genetic testing practices and judge our success against a validation cohort born 2020-2021. We cannot exclude pandemic effects leading to important but otherwise occult differences between the training/test and validation cohorts, though, the model performance in the train/test and validation cohorts was similar. In a similar vein, the ideal target outcome is clinical utility or usefulness of genetic testing to one or more stakeholders. The recent explosion in utility measures for genome sequencing and/or genetic medicine services in ICUs had not happened when the vast majority of training data for this study was generated. It is logical that the next generation of predictive analytics in this space will use utility as a target variable, but until 1000s of such measures can accrue, it seems a reasonable assumption that most genetic testing in patients admitted to or decently discharged from a Level IV NICU is ordered because someone thinks it would be helpful.

Estimation of benefit from simulated recommendations

The time at which classifiers made predictions had minimal impact on their performance (ROC and PR AUC scores increased by 0.01) but modelling this allowed for a more robust simulation of possible effects of implementation. Importantly, many of the positive predictions could be made as early as day 1 in the Level IV NICU, demonstrating that accrual of phenotypic information relevant to making predictions of genetic testing need occurs early in the NICU course.

In our simulations, using ML as a basis for genetic testing policies decisively impacted the overall length of diagnostic odyssey as well as the number of subjects able to complete their odysseys within 14 days of Level IV NICU admission, independent of any change in test selection. However, we also see that the positive effects of ML recommendations are enhanced when such recommendations are accompanied by a policy of rGS as a first line test.

The reduction in diagnostic odyssey was most pronounced in patients who eventually received GS. The distribution of estimated benefit in the subpopulation of patients who received GS was highly

bimodal, with many patients who did receive early rapid testing and did not directly benefit from the model's recommendation, as well many others who received GS much later and experienced very large individual benefit. This is likely attributable to historical usage of ES/GS as a last-line test, ordered after many other options have already been pursued, causing this population of patients to have the longest diagnostic odysseys. As early-in-life GS (and especially rGS) becomes more common, this effect should diminish.

As we saw, the ML model made positive predictions for 280 patients who had genetic testing but not GS. Durations of diagnostic odysseys for this group are censored in the present study, but are known in the literature to persist for years[33]. By our estimation, rGS can make the same diagnoses as other testing so could replace all their other testing (with the caveat that a small number of edge cases like methylation disorders may not be detected by rGS). Though rGS is more expensive on an individual test basis than the other tests, early rGS can cut down on the time and money spent on ordering and obtaining results for multiple, less comprehensive genetic tests. Additionally, some of these patients would likely derive greater utility from genome sequencing than their actually ordered test, either because GS could make a diagnosis or because non-diagnostic GS may offer more utility than a non-diagnostic NGS panel in some cases.

Patients with a known genetic diagnosis saw smaller reductions in diagnostic odyssey. Definitionally, patients in this group have reached a conclusive end to their diagnostic odyssey. By comparison, many patients without a diagnosis have not yet seen the end of their diagnostic odyssey, meaning that group will necessarily have longer and more extreme diagnostic odysseys. The full extent of each patient's diagnostic odyssey is difficult to capture. Last recorded test was intended as a conservative measure of the most recent time at which a patient suspected of genetic illness was still being evaluated. However, a lack of follow-up testing – especially for patients who never got as far as genomic sequencing – may not signify the end of the diagnostic odyssey. Additionally, some patients are lost to follow-up due to death, family relocation, or other circumstances which are not discernable by our data acquisition strategy. Conceivably, patients with no positive genetic test as of June 2023 may still be undergoing genetic evaluation. In these cases, our estimates of diagnostic odyssey length and benefit should be under-estimates.

Patients diagnosed with classical aneuploidies were almost all (19/20) detected by the model. These patients, predictably, received little or no benefit from an rGS recommendation as their diagnoses had already been determined before the recommendation timepoint in all but one case. However, it is desirable that the model identify these patients rather than ignore them, in the improbable event that such a patient does not receive a diagnosis as early as they should. In general, we see from Figure 5 that many patients with shorter diagnostic odysseys are recovered by the model and patients with some of the longest odysseys are the most frequently missed by the model. In essence, the model offers some benefits to the cases that human experts struggle with, while replicating expert recognition in the easiest cases.

Because of the retrospective nature of this study, we could not assess the true impact of any human-machine interactions once recommendations are made. It is possible that recommendations on genetic testing by the machine learning algorithm may be equally accurate across the first several days in the level IV NICU, while the willingness of the clinician to pursue broad genetic testing is higher on the second or third day as other diagnostic studies (e.g. blood cultures) begin to result. Similarly, without prospective evaluation it is not possible to know what the effect on clinician behavior of negative recommendations is (i.e. if clinicians trust the model predictions will they refrain from genetic testing when their clinical intuition conflicts with the model's recommendations?).

Additionally, we have ascribed zero benefit to cases where the model recommended testing patients identified as having "no genetic involvement" and regarded these as extraneous recommendations. However, these patients may have had testing outside of our healthcare system, or simply failed to have been recognized for testing. We also, via chart review, document many cases that likely would be considered to derive benefit under constantly evolving standards for use of rGS. Additionally, given the evidence that clinicians find negative rGS results to be useful, it is plausible that at least some of these "false positive" recommendations would actually be valuable to neonatologists[34].

Future directions and expansions

While we used data from Level III NICUs in training our model, we developed it with the goal of making predictions in Level IV NICUs. It is tempting to want to extrapolate our findings and/or approach to Level III NICUs which care for many more patients than Level IV NICUs. There are several challenges to developing a model under this approach in the setting of the level III NICU, some of which are different than the challenges in developing a model in the Level IV setting. We expect a higher prevalence of genetic illness in the level IV NICU and hence more balanced data and better overall classifier performance. In the Level III setting, one would have to make careful consideration as to how to handle patients who would ultimately move to level IV. Different exclusion and labelling criteria may enable us to adapt this approach to the Level III NICU setting, but the current model should not be used for this purpose.

The phenotypes for the training data were obtained from clinical text by ClinPhen. As any NLP tool is likely to have its own idiosyncrasies, this may create exotic or artifactual behaviors in the model's predictions and may hinder model portability. Moreover, it can be computationally costly and administratively difficult to obtain clinical text from the health record and apply NLP tools to extract phenotype concepts from this text. This may hinder model portability. An alternative would be to mine the widely used ICD codes from the structured health record and map them to phecodes or pediatric phecodes[35,36] to create an alternative encoding of the patient's phenotype. This could make for a more portable model.

Features were included in this model to account for degree of prematurity, as isolated prematurity is a common non-genetic reason for admission to the NICU. Infection is another common non-genetic reason for NICU admission; the existence and results of blood cultures and other common tests for

infectious disease could be leveraged to improve the model's ability to distinguish cases in which the patient's phenotype can be explained by infection.

Additional features accounting for the specialties of providers who have encountered a patient could help the model recognize patterns of complex phenotypes requiring the attention of a genetics provider. In a way, the current model (suboptimally) encodes some of this information by conflating Hematology-Oncology consultation with the HPO term for Neoplasms. Appropriately engineering this feature may secondarily correct this behavior.

The potential benefit of this model was estimated retrospectively and under the assumption of 100% adherence to simulated policies. In reality, clinical decision support systems seldom have such high adherence. A more meaningful evaluation of this model's benefit would need to be carried out in the prospective setting of a randomized control trial. In this setting, utility of the test recommendations could be assessed by C-Guide[37] or other survey instrument to determine the perceived utility of the testing recommended by the model.

In summary, in a retrospective analysis using a holdout validation sample with follow-up through 18 months, NeoGx accurately predicts need for genetic testing in patients admitted to a Level IV NICU. In simulation of multiple decision-making policies, CDS based on NeoGx shortens diagnostic odysseys, moreso when combined with broadened adoption of rGS. Failure mode analysis identifies areas for improvement, but the most significant progress to be made is in the areas of prospective validation, adoption of clinical utility as an outcome measure and/or generalization beyond a single center.

Data availability

All data produced in the present study are available upon reasonable request to the authors.

Acknowledgements

The authors would like to acknowledge the work of Rajesh Ganta, Brent Merryman, and Nan Zhang, analysts who obtained data for this study from the NCD RDW and from clinical data repositories. Steve Rust and Sven Bambach on our research institute's Data Science team were consulted on machine learning study design and on approaches to bias evaluation.

Funding

BPCs work on this publication was supported, in part, by the National Center for Advancing Translational Sciences of the National Institutes of Health under Grant Number **UM1TR004548.** The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

Conceptualization: A.A.A., B.P.C.; Data curation: R.M., M.M.; Formal analysis: A.A.A.; Funding acquisition: B.P.C.; Investigation: A.A.A., B.P.C.; Methodology: A.A.A., B.P.C.; Project administration: B.P.C.; Resources: B.P.C.; Software: A.A.A.; Supervision: B.P.C; Validation: A.A.A.; Visualization: A.A.A.; B.P.C.; Writing-original draft: A.A.A; Writing-review & editing: A.A.A., B.P.C.

Ethics Declaration

The Abigail Wexner Research Institute IRB approved the study (STUDY00000276) with a waiver of informed consent.

Conflicts of Interest

The authors have no conflicts to disclose.

ABBREVIATIONS

NICU: Neonatal Intensive Care Unit

NCH: Nationwide Children's Hospital

EHR: Electronic Health Record

RDW: Research Data Warehouse

Dx: Diagnosis ("received a Dx") or Diagnostic ("Dx odyssey")

rGS: rapid Genomic Sequencing

GS: Genome Sequencing

ES: Exome Sequencing

NGS: Next Generation Sequencing

H&P: History and Physical Examination

GA: Gestational Age

BW: Birth Weight

HPO: Human Phenotype Ontology

IC: Information Content

IPR: Information Potential Ratio

ML: Machine Learning

CI: Confidence Interval

SHAP: SHapley Additive explanation

ROC: Receiver Operator Characteristic

PR: Precision-Recall

AUC: Area Under the Curve

REFERENCES

1. Marouane A, Olde Keizer R a. CM, Frederix GWJ, Vissers LELM, de Boode WP, van Zelst-Stams W a. G. Congenital anomalies and genetic disorders in neonates and infants: a single-center observational cohort study. *Eur J Pediatr*. 2022;181(1):359-367. doi:10.1007/s00431-021-04213-w

2. Weiner J, Sharma J, Lantos J, Kilbride H. How Infants Die in the Neonatal Intensive Care Unit: Trends From 1999 Through 2008. *Arch Pediatr Adolesc Med*. 2011;165(7):630-634. doi:10.1001/archpediatrics.2011.102

3. Simeoni U. How infants die in neonatal intensive care units - a European perspective. *Acta Paediatr Oslo Nor 1992*. 2012;101(6):552-554. doi:10.1111/j.1651-2227.2012.02685.x

4. Synnes AR, Berry M, Jones H, et al. Infants with congenital anomalies admitted to neonatal intensive care units. *Am J Perinatol*. 2004;21(4):199-207. doi:10.1055/s-2004-828604

5. Improved National Prevalence Estimates for 18 Selected Major Birth Defects—United States, 1999-2001. *JAMA*. 2006;295(6):618-620. doi:10.1001/jama.295.6.618

6. Parker SE, Mai CT, Canfield MA, et al. Updated National Birth Prevalence estimates for selected birth defects in the United States, 2004-2006. *Birt Defects Res A Clin Mol Teratol*. 2010;88(12):1008-1016. doi:10.1002/bdra.20735

7. Stallings EB, Isenburg JL, Rutkowski RE, et al. National population-based estimates for major birth defects, 2016–2020. *Birth Defects Res*. 2024;116(1):e2301. doi:10.1002/bdr2.2301

8. Ferreira CR. The burden of rare diseases. *Am J Med Genet A*. 2019;179(6):885-892. doi:10.1002/ajmg.a.61124

9. Haendel M, Vasilevsky N, Unni D, et al. How many rare diseases are there? *Nat Rev Drug Discov*. 2020;19(2):77-78. doi:10.1038/d41573-019-00180-y

10. Messick EA, Backes CH, Jackson K, Conroy S, Hart SA, Cua CL. Morbidity and mortality in neonates with Down Syndrome based on gestational age. *J Perinatol Off J Calif Perinat Assoc*. 2023;43(4):445-451. doi:10.1038/s41372-022-01514-2

11. Swaggart KA, Swarr DT, Tolusso LK, He H, Dawson DB, Suhrie KR. Making a Genetic Diagnosis in a Level IV Neonatal Intensive Care Unit Population: Who, When, How, and at What Cost? *J Pediatr*. 2019;213:211-217.e4. doi:10.1016/j.jpeds.2019.05.054

12. Callahan KP, Clayton EW, Lemke AA, et al. Ethical and Legal Issues Surrounding Genetic Testing in the NICU. *NeoReviews*. 2024;25(3):e127-e138. doi:10.1542/neo.25-3-e127

13. Authors, Basharat S, Smith A, Darvesh N, Rader T. *2023 Watch List: Top 10 Precision Medicine Technologies and Issues: CADTH Horizon Scan*. Canadian Agency for Drugs and Technologies in Health; 2023. Accessed May 30, 2024. http://www.ncbi.nlm.nih.gov/books/NBK596300/

14. Petrikin JE, Willig LK, Smith LD, Kingsmore SF. Rapid whole genome sequencing and precision neonatology. *Semin Perinatol*. 2015;39(8):623-631. doi:10.1053/j.semperi.2015.09.009

15. Saunders CJ, Miller NA, Soden SE, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med*. 2012;4(154):154ra135. doi:10.1126/scitranslmed.3004041

16. Smith HS, Ferket BS, Gelb BD, et al. Parent-Reported Clinical Utility of Pediatric Genomic Sequencing. *Pediatrics*. 2023;152(2):e2022060318. doi:10.1542/peds.2022-060318

17. Dimmock D, Caylor S, Waldman B, et al. Project Baby Bear: Rapid precision care incorporating rWGS in 5 California children's hospitals demonstrates improved clinical outcomes and reduced costs of care. *Am J Hum Genet*. 2021;108(7):1231-1238. doi:10.1016/j.ajhg.2021.05.008

18. Maron JL, Kingsmore S, Gelb BD, et al. Rapid Whole-Genomic Sequencing and a Targeted Neonatal Gene Panel in Infants With a Suspected Genetic Disorder. *JAMA*. 2023;330(2):161-169. doi:10.1001/jama.2023.9350

19. Wojcik MH, Del Rosario MC, Agrawal PB. Perspectives of United States neonatologists on genetic testing practices. *Genet Med Off J Am Coll Med Genet*. 2022;24(6):1372-1377. doi:10.1016/j.gim.2022.02.009

20. Seither K, Thompson W, Suhrie K. A Practical Guide to Whole Genome Sequencing in the NICU. *NeoReviews*. 2024;25(3):e139-e150. doi:10.1542/neo.25-3-e139

21. Wojcik MH, Callahan KP, Antoniou A, et al. Provision and availability of genomic medicine services in Level IV neonatal intensive care units. *Genet Med Off J Am Coll Med Genet*. 2023;25(10):100926. doi:10.1016/j.gim.2023.100926

22. Morley TJ, Han L, Castro VM, et al. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat Med*. 2021;27(6):1097-1104. doi:10.1038/s41591-021-01356-z

23. Automated prioritization of sick newborns for whole genome sequencing using clinical natural language processing and machine learning - PubMed. Accessed April 9, 2024. https://pubmed.ncbi.nlm.nih.gov/36927505/

24. Deisseroth CA, Birgmeier J, Bodle EE, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med Off J Am Coll Med Genet*. 2019;21(7):1585-1593. doi:10.1038/s41436-018-0381-1

25. The Human Phenotype Ontology in 2021 - PubMed. Accessed April 9, 2024. https://pubmed.ncbi.nlm.nih.gov/33264411/

26. Chou JH, Roumiantsev S, Singh R. PediTools Electronic Growth Chart Calculators: Applications in Clinical Care, Research, and Quality Improvement. *J Med Internet Res*. 2020;22(1):e16204. doi:10.2196/16204

27. Jagadeesh KA, Birgmeier J, Guturu H, et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet Med Off J Am Coll Med Genet*. 2019;21(2):464-470. doi:10.1038/s41436-018-0072-y

28. Boyd K, Santos Costa V, Davis J, Page CD. Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation. *Proc Int Conf Mach Learn Int Conf Mach Learn*. 2012;2012:349.

29. Lunke S, Bouffler SE, Patel CV, et al. Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nat Med*. 2023;29(7):1681-1691. doi:10.1038/s41591-023-02401-9

30. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv*. 2021;54(6):115:1-115:35. doi:10.1145/3457607

31. Gubbels CS, VanNoy GE, Madden JA, et al. Prospective, phenotype-driven selection of critically ill neonates for rapid exome sequencing is associated with high diagnostic yield. *Genet Med Off J Am Coll Med Genet*. 2020;22(4):736-744. doi:10.1038/s41436-019-0708-6

32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc.; 2017:4768-4777.

33. Wu AC, McMahon P, Lu C. Ending the Diagnostic Odyssey: Is whole genome sequencing the answer? *JAMA Pediatr*. 2020;174(9):821-822. doi:10.1001/jamapediatrics.2020.1522

34. Dimmock DP, Clark MM, Gaughran M, et al. An RCT of Rapid Genomic Sequencing among Seriously Ill Infants Results in High Clinical Utility, Changes in Management, and Low Perceived Harm. *Am J Hum Genet*. 2020;107(5):942-952. doi:10.1016/j.ajhg.2020.10.003

35. Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci*. 2021;4:1-19. doi:10.1146/annurev-biodatasci-122320-112352

36. Grabowska ME, Van Driest SL, Robinson JR, et al. Developing and evaluating pediatric phecodes (Peds-Phecodes) for high-throughput phenotyping using electronic health records. *J Am Med Inform Assoc JAMIA*. 2024;31(2):386-395. doi:10.1093/jamia/ocad233

37. Hayeems RZ, Luca S, Ungar WJ, et al. The Clinician-reported Genetic testing Utility InDEx (C-GUIDE): Preliminary evidence of validity and reliability. *Genet Med Off J Am Coll Med Genet*. 2022;24(2):430-438. doi:10.1016/j.gim.2021.10.005