Genome Medicine

RESEARCH Open Access

Automated prioritization of sick newborns for whole genome sequencing using clinical natural language processing and machine learning

Bennet Peterson[1], Edgar Javier Hernandez[2], Charlotte Hobbs[3], Sabrina Malone Jenkins[4], Barry Moore[2], Edwin Rosales[3], Samuel Zoucha[4], Erica Sanford[3,5], Matthew N. Bainbridge[3], Erwin Frise[6], Albert Oriol[7], Luca Brunelli[4], Stephen F. Kingsmore[3] and Mark Yandell[2*]

Abstract

Background Rapidly and efficiently identifying critically ill infants for whole genome sequencing (WGS) is a costly and challenging task currently performed by scarce, highly trained experts and is a major bottleneck for application of WGS in the NICU. There is a dire need for automated means to prioritize patients for WGS.

Methods Institutional databases of electronic health records (EHRs) are logical starting points for identifying patients with undiagnosed Mendelian diseases. We have developed automated means to prioritize patients for rapid and whole genome sequencing (rWGS and WGS) directly from clinical notes. Our approach combines a clinical natural language processing (CNLP) workflow with a machine learning-based prioritization tool named *Mendelian Phenotype Search Engine* (MPSE).

Results MPSE accurately and robustly identified NICU patients selected for WGS by clinical experts from Rady Children's Hospital in San Diego (AUC 0.86) and the University of Utah (AUC 0.85). In addition to effectively identifying patients for WGS, MPSE scores also strongly prioritize diagnostic cases over non-diagnostic cases, with projected diagnostic yields exceeding 50% throughout the first and second quartiles of score-ranked patients.

Conclusions Our results indicate that an automated pipeline for selecting acutely ill infants in neonatal intensive care units (NICU) for WGS can meet or exceed diagnostic yields obtained through current selection procedures, which require time-consuming manual review of clinical notes and histories by specialized personnel.

*Correspondence:
Mark Yandell
myandell@genetics.utah.edu
Full list of author information is available at the end of the article

Background

It is estimated that 7 million infants are born worldwide with genetic disorders each year [1]. Admission to the neonatal intensive care unit (NICU) often provides the first opportunity for their diagnosis and treatment. Disease can progress rapidly in acutely ill infants, necessitating timely diagnosis in the hope of implementing personalized interventions that can decrease morbidity and mortality. Thus, rapid whole genome sequencing (rWGS) is increasingly being used as a first line diagnostic test [2, 3].

Current estimates suggest that around 18% of neonates admitted to the NICU harbor a Mendelian disease, and rWGS diagnostic rates in this population are over 35% [4, 5]. Rapidly and efficiently identifying infants for WGS is costly and challenging, as large NICUs often see more than 1000 admissions per year, and neonatal clinical histories evolve rapidly from the time of admission. Previous studies of rWGS in the NICU used inclusion criteria that limited enrollment to the first 96 h [3, 5] or 7 days [6] of admission or development of an abnormal response to standard therapy for an underlying condition, but these restrictions may miss the earliest opportunity to sequence a neonate. Minute-to-minute changes in laboratory results, diagnostic imaging, and clinical trajectory suggest that constant automated vigilance, as opposed to one or two isolated points in time, may be optimal to identify infants most likely to benefit from WGS. Done manually, this would be prohibitively time-consuming and costly. Automated means to prioritize patients for WGS are thus badly needed. Indeed, this is the principal motivation for the work described here.

Phenotype descriptions are crucial components of the WGS diagnostic process, and many tools exist for combining phenotypic terms with WGS data to prioritize disease-causing variants [7–10]. Current best practice is to describe patient phenotypes using Human Phenotype Ontology (HPO) terms [11]. These descriptions usually take the form of machine-readable phenotype term lists, an important prerequisite for automated analyses.

Care providers emphasize the importance of clinical notes for informing disease diagnosis, and HPO-based phenotype descriptions are generally compiled through manual review of these free text documents. Unfortunately, this is a time-consuming process that requires highly trained experts and is a major bottleneck for application of WGS in the NICU [12, 13].

Natural language processing (NLP) is a class of computational methods for generating structured data from unstructured free text. Recent work has begun to explore the utility of using clinical natural language processing technologies (CNLP) to automatically generate descriptions directly from clinical notes, with several groups demonstrating that rWGS diagnosis rates using CNLP derived descriptions can equal or exceed those obtained using manually compiled ones [12, 14]. This is a significant step towards scalability and automation. The ability to automatically survey all NICU admissions daily, for example, would mean that rWGS candidates could be ranked as part of an ever-evolving triage process based upon the latest contents of their EHRs.

Although the use of HPO descriptions for WGS-based Mendelian diagnosis is now established practice [7–10, 14], the benefit of prioritization of patients for sequencing based on HPO terms is not known. To explore the feasibility of such an approach, we have combined a CNLP workflow with a machine learning-based prioritization tool we call the Mendelian Phenotype Search Engine (MPSE) [15]. MPSE employs HPO-based phenotype descriptions derived from patient EHRs to compute a score. This score can be used to determine the likelihood that a Mendelian condition is contributing to a patient's clinical presentation and, thus, can be used for the prioritization of patients for WGS. To demonstrate feasibility, we used a highly curated clinical dataset consisting of 1049 patients admitted to a level IV NICU (the highest level of acuity for a NICU) and their clinic notes; 293 of these children had rWGS, with 85 receiving a diagnosis. Our cross validated results indicate that an entirely automated CNLP/MPSE-based selection process for rWGS can obtain diagnostic rates equaling or exceeding those obtained though manual review and selection as per current best practice. A second independent replication study at the University of Utah provides additional support for these conclusions, demonstrating that MPSE operates effectively at both institutions.

Methods

Datasets

Our clinical dataset consisted of 293 probands who underwent rWGS at Rady Children's Hospital in San Diego (RCHSD), 85 of which received a molecular diagnosis of Mendelian disorder. These cases were a sample of convenience drawn from symptomatic children enrolled in previously published studies that examined the diagnostic rate, time to diagnosis, clinical utility, outcomes, and health care utilization of rWGS between 26 July 2016 and 25 September 2018 at RCHSD (ClinicalTrials.gov identifiers: NCT03211039, NCT02917460, and NCT03385876) [2, 5, 12, 16, 17]. All subjects had a symptomatic illness of unknown etiology in which a genetic disorder was suspected. The diagnosed individuals represent a real-world population comprised of different Mendelian conditions resulting from diverse modes of disease inheritance and disease-causing genotypes [3, 14]. To this cohort, we added every NICU admission at RCHSD in the year 2018. The 756 additional patients and their clinic notes provide a diversity of phenotypes not necessarily associated with Mendelian diseases. In total, the RCHSD dataset consisted of 1049 individuals.

A second independent dataset of 35 probands that were sequenced as part of the University of Utah NeoSeq program [18] and 2930 randomly selected (as per IRB; see Declarations) University of Utah level III NICU patients from 2010 to 2022 was retrospectively analyzed to evaluate the utility of the RCHSD training data for prioritizing probands for rWGS at a second institution. Additional file 1: Table S1 and Table S2 show clinical diagnosis frequencies for sequenced RCHSD and Utah NeoSeq cases broken down by positive/negative rWGS diagnostic status. These tables highlight the variety and complexity of Mendelian disease phenotypes found in upper level NICUs. They also show a lack of overrepresented disease and phenotype categories among cases or controls. This lack of recurrent signal is consistent with the fact that there are over 7000 known Mendelian diseases, many of which have highly variable phenotypes. These facts led us to pursue a general, rather than disease-by-disease approach for prioritizing probands for rWGS.

Phenotype descriptions

Highly curated, manually created HPO-based phenotype descriptions were provided for each of the 293 RCHSD and 35 University of Utah WGS cases, as described in NSIGHT1 [3]. Corresponding CNLP-derived phenotype descriptions were generated for all 1049 RCHSD probands and 2965 University of Utah probands by NLP analysis of clinical notes using CLiX ENRICH (Clinithink, Alpharetta, GA) [14, 19]. Clinical notes dated post-rWGS were excluded from analysis to prevent possible confounding from knowledge of sequencing results. CLiX was run in default mode with “acronyms on.”

MPSE

The Mendelian phenotype Search Engine (MPSE) employs Human Phenotype Ontology (HPO)-based descriptions to prioritize patients, determining the likelihood that a Mendelian condition underlies a patient’s phenotype, based upon a training dataset. MPSE does not attempt to determine which Mendelian disease might underlie the patient’s phenotype, rather it seeks to categorize patients as positive or negative for Mendelian disease. MPSE employs a simple, well-established approach: Naïve Bayes [20]. Briefy, MPSE uses the differences in HPO term frequencies between a collection of cases and controls to score each proband. The algorithm employs the BernoulliNB package from *scikit-learn*, a general-purpose machine learning library written in the Python programming language [21]. We also discovered that the number of terms in a proband’s HPO description correlated modestly with age ($r^2 = 0.0725$); accordingly, we used a linear regression to control for this effect. Although one can envision many algorithmic approaches to classification other than Naïve Bayes, e.g., support vector machines or neural nets, for this proof of principle study, we sought to demonstrate feasibility and provide baseline performance metrics. Future work will explore more sophisticated approaches to data modeling.

Cross validation

We validated our results using leave-one-out cross validation (i.e., k-fold cross validation, with $k = 1$) [22]. More specifically, using the RCHSD data, we created 1049 different training datasets—each differing by a single proband—scoring each proband against a (different) version of MPSE, trained using a data subset that did not contain the proband being scored. All performance metrics were computed using this cross-validation scheme. Using the cross-validated model derived from the RCHSD dataset, we then carried out an independent replication study using the clinical notes of 2965 University of Utah level III NICU admits. This dataset includes 35 WGS probands sequenced to date by the University of Utah NeoSeq program [18].

Results and discussion

Previous work [10, 12, 23], including our own [14], has demonstrated the utility of HPO-based, CNLP-derived phenotype descriptions for post sequencing diagnostic applications. Here, we explore the feasibility of using CNLP phenotype descriptions, manufactured using the same NLP protocols, for triaging patients for WGS. To do so, we combined a natural language processing (NLP) workflow based around the commercially available CLiX tool [19] with an ML-based prioritization tool we call MPSE, the Mendelian Phenotype Search Engine.

MPSE (see the “Methods” section) employs the Human Phenotype Ontology (HPO) [11] to prioritize patients. The a priori likelihood that a patient has a Mendelian condition is a computed probability based on the existence of HPO terms in the patient’s phenotype that are similar to those patients who previously had WGS. To investigate feasibility, we utilized curated RCHSD clinical data: 1049 level IV NICU admissions and their clinical notes. Of these 1049 patients, 293 had rWGS and 85 received a molecular diagnosis. We validated the results presented below using leave-one-out cross validation; see the “Methods” section for details. To examine the broader applicability of the RCHSD training data to other NICUs, we also carried out a second independent

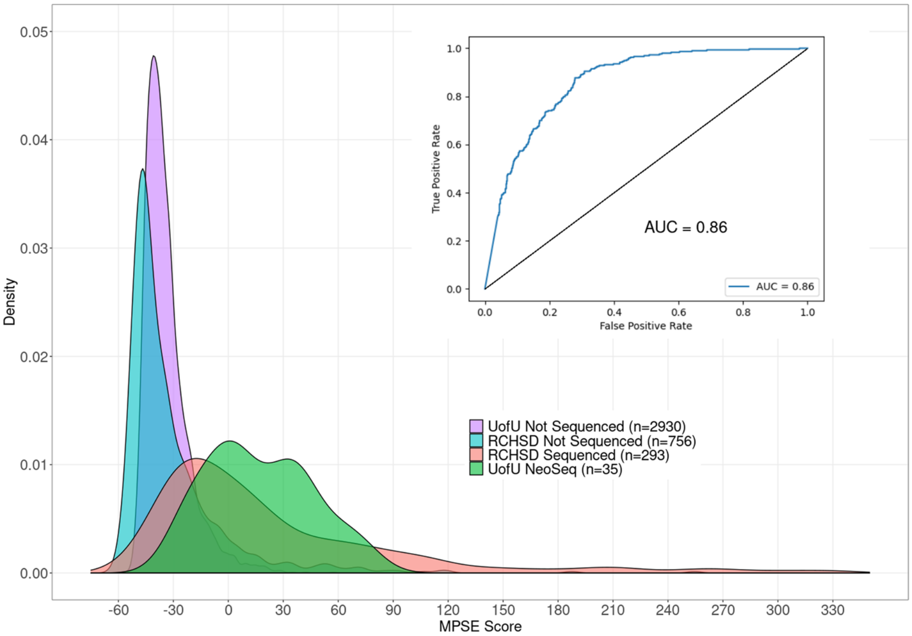

Fig. 1 Automatically identifying probands with Mendelian phenotypes and prioritizing them for WGS using NLP-derived HPO phenotype descriptions. Distributions of MPSE raw scores for RCHSD sequenced (red) and RCHSD unsequenced (blue) probands. Score distributions for Utah NeoSeq (green) and Utah unsequenced probands (purple). Insert: Receiver operator characteristic (ROC) curve for RCHSD data. MPSE scores are -log likelihood ratios

replication study using the clinical notes of 2965 patients from the University of Utah level III NICU.

Automated generation of HPO terms

We obtained HPO phenotype descriptions for all probands from clinical notes using Clinithink, a third party NLP tool [19]. Automatically generating phenotypic descriptions via NLP is a major strength, as it enables the creation of large and dynamic pools of HPO-based phenotype descriptions for downstream prioritization activities.

Comparison of the CNLP descriptions to their corresponding manually compiled ones revealed notable differences with regards to HPO term numbers and contents. The CLiX generated descriptions for the RCHSD and NeoSeq cohorts had an average of 114.8 terms (min: 3, median: 91, max: 1000) and 64.5 terms (min: 1, median: 58, max: 300) respectively, whereas the corresponding manually created descriptions averaged 4.1 terms (min: 1, median: 3, max: 24) and 9.5 terms (min: 3, median: 9, max: 16) respectively.

Prioritizing patients

We first sought to evaluate how effective our CNLP/MPSE pipeline was at prioritizing patients for WGS. In other words, did the children originally selected for WGS by physicians have higher MPSE scores than those who were not selected? Figure 1 demonstrates that this is the case. As can be seen, the distributions of MPSE raw scores for the RCHSD and Utah WGS-selected children are well-separated from unsequenced ones. RCHSD sequenced cases had an average MPSE score of 26.6 while unsequenced controls had an average score of -31.7, statistically different by Student's independent samples *t*-test ($p < 2\text{e}{-16}$). The difference in mean MPSE score between Utah sequenced cases (17.3) and unsequenced controls (-33.7) was also statistically different ($p = 2\text{e}{-12}$). The insert shows a receiver operator characteristic (ROC) curve for the RCHSD data (AUC 0.86), indicating that MPSE can effectively prioritize probands for rWGS. The corresponding AUC for the Utah data was 0.85, essentially identical to the RCHSD result (ROC curve not shown). A possible clinical application scenario can be imagined where MPSE score cutoffs are used to prioritize patients for further review by physicians. For the RCHSD training cohort described here, for example, taking only MPSE scores > 30 would prioritize 30% (89/293) of cases and 4% (31/756) of controls, while taking only MPSE scores > 90 would prioritize 14% (40/293) of cases and 0.8% (6/756) of controls. Anonymized MPSE scores for each patient in these cohorts are tabulated in Additional file 1: Tables S3 and S4.

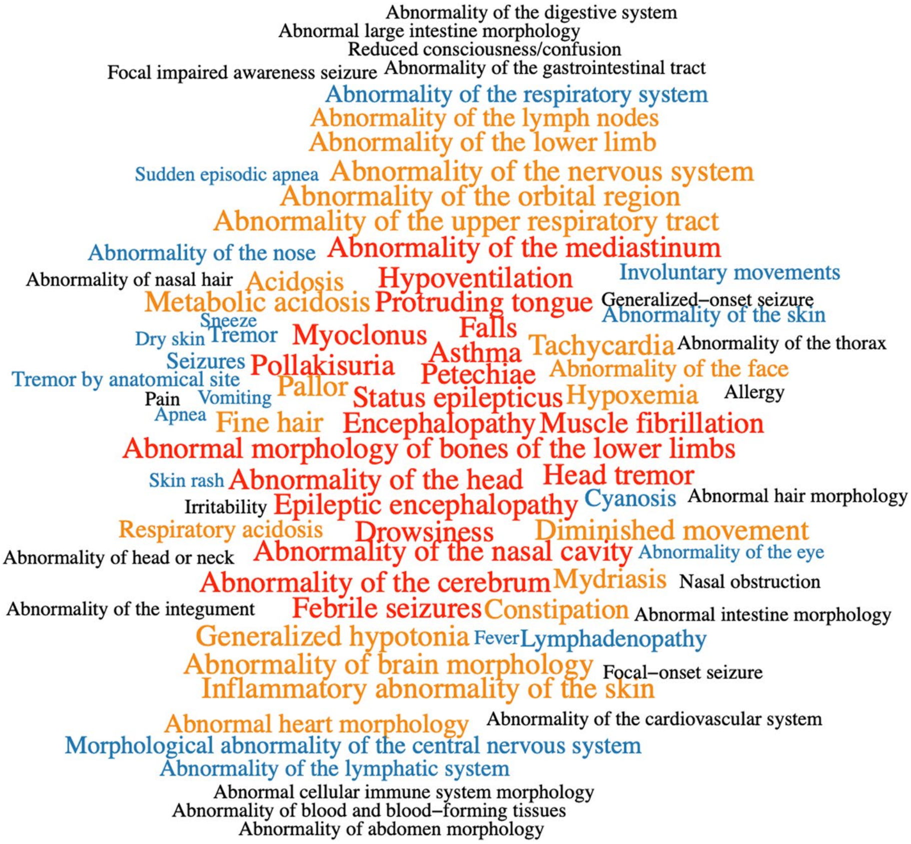

Fig. 2 An automatically generated HPO-based phenotype description scored by MPSE. In this word-cloud, size and color are proportional to each HPO term's contribution to the proband's final MPSE prioritization score. Previously diagnosed by RCHSD using WGS, this child is heterozygous for a large deletion on the X chromosome which spans the PCDH19 gene, causative for female-restricted X-linked epileptic encephalopathy

Cardinal phenotype terms

MPSE also provides means to identify, and highlight for expert review, those terms in a phenotype description that are most consistent with Mendelian disease. We refer to these terms as the proband's cardinal phenotypes. Figure 2 shows a CNLP phenotype description as a word cloud, wherein font sizes have been scaled by their individual contributions to the proband's final MPSE score; those with the highest scores are shown in red; these are the proband's MPSE cardinal phenotypes. These views of the patient's phenotype description are designed to speed physician review and improve explainability.

MPSE diagnostic rates

To estimate MPSE-driven diagnostic rates, RCHSD and University of Utah sequenced probands were scored using leave-one-out cross validation, as described in the "Methods" section. The diagnostic fraction for these cohorts was 29% (85/293) and 43% (15/35), respectively. It should be borne in mind that this RCHSD diagnostic rate is for the specific dataset under analysis. It is not the RCHSD institutional WGS diagnostic rate. To facilitate comparison between these groups, we randomly re-sampled the larger RCHSD dataset so that it too had a 43% (85/198) diagnostic rate.

Figure 3 shows projected diagnostic rates for these cohorts as a function of their MPSE scores. The negative slopes of the red, green, and blue curves indicate that when using CNLP, higher MPSE scores are associated with diagnosed probands at both institutions. For instance, the top 25% of probands ranked on their MPSE scores from CNLP-generated phenotypes show very high diagnostic rates, approaching 100% for the highest MPSE scores. Moreover, for the CNLP datasets, diagnostic rates remain at or above the cohort diagnostic fraction of 43% at every MPSE score percentile. In contrast, the MPSE scores calculated from manually curated phenotypes (gray curve) are at best weakly associated with diagnostic status. This is not a result of inferiority of the physician-generated phenotypes;

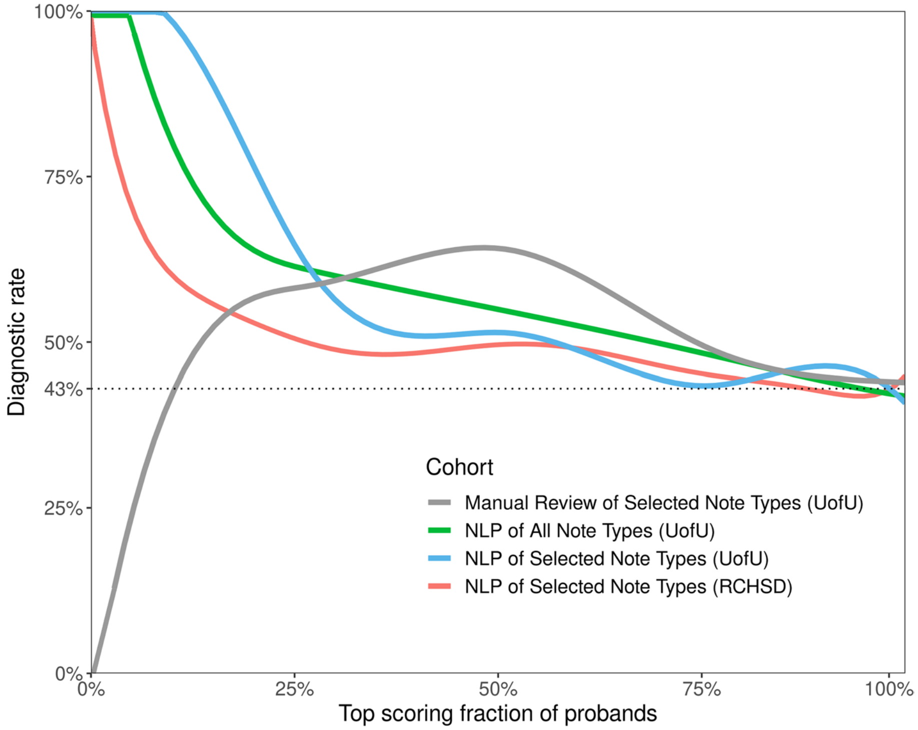

Fig. 3 MPSE projected diagnostic rates. Higher MPSE scores correspond to increased probability of diagnosis, and projected diagnostic rates remain at or above the cohort diagnostic fraction of 43% at every MPSE score percentile

rather, it is due to the fact that MPSE was trained using deep CNLP-derived phenotype data; recall that CNLP compared to manual review resulted in 64.5 vs 9.5 HPO terms/proband, respectively. Collectively, these results indicate that an MPSE-based prioritization pipeline in conjunction with manual review could increase diagnostic rates above those obtained solely through expert manual case-review.

Impact of note types

Both RCHSD and the University of Utah limit manual review of clinical notes to a subset of note types deemed most informative by their institution's expert reviewers. This is done to speed review by avoiding less informative and redundant note types. A potential advantage of CNLP is that volume is no longer an issue, and every note can be processed. We thus sought to evaluate the utility of processing all notes for every proband. The results of this experiment are also shown in Fig. 3, where the blue and green curves summarize diagnostic enrichment as a function of MPSE score and note volumes. AUC for the top 50% of high scoring probands using all clinical notes vs. using only the selected note types is quite similar—62% and 65%, respectively. Thus, for the Utah dataset, using all available notes for every proband does not negatively impact diagnostic rates.

Impact of patient populations

It is worth noting that underlying NICU populations differ between RCHSD and the University of Utah. Whereas RCHSD is a level IV NICU, the University of Utah operates a level III NICU, with the most severely ill patients transferred to Intermountain's Primary Children's neighboring level IV facility. Thus, patients in the Utah dataset are likely to have fewer conditions requiring surgical interventions and a higher level of intensive care. Despite being trained using the RCHSD level IV data, Fig. 1 makes it clear that the lesser acuity of level III patients compared to level IV patients did not interfere with MPSE's ability to identify suitable candidates for sequencing nor did it negatively impact the correlation between MPSE score and Mendelian diagnostic rates (Fig. 3). This finding suggests MPSE's robustness to differences in NICU patient populations.

Conclusions

We have demonstrated the feasibility of prioritizing individuals for WGS, using automated means, and that supplementing clinical review with this automated process could meet or exceed diagnostic yields obtained solely through manual review of clinical notes. More sophisticated machine learning techniques might further improve the accuracy of prioritization.

Neural and Bayesian networks and random forest-based approaches generally outperform naïve Bayes. Likewise, addition of other metadata such as provider billing codes, medication histories, ancestry, and socio-economic indicators might still further improve performance. Nevertheless, even without such enhancements, our CNLP/MPSE workflow prioritized patients for rWGS with relatively high accuracy (AUC = 0.86), with maximal projected diagnostic yields highly enriched for the top scoring quartile. These results bode well for future improved versions of the pipeline.

The ability of MPSE to accurately distinguish sequenced from unsequenced probands at both RCHSD and the University of Utah demonstrates the generalizability of the RCHSD training data, at least between two leading research institutions. The fact that MPSE was trained using RCHSD's level IV NICU patients and replicated in Utah's level III NICU also provides some indication of MPSE's robustness and applicability. Broader generalization, however, remains to be proven. Generalization is important because as WGS-based diagnosis becomes more widespread, and patients considered for testing become more diverse, clinical cultures and institutional differences in clinical note taking might render the parameters derived from the RCHSD training dataset less effective at some sites. In this regard, the ability of the pipeline to consume all notes for every proband is clearly an advantage, as it means adopters need not establish cross institutional equivalents in note types; instead, they can simply harvest every available clinical note for every proband.

More broadly, generalizability of training data must be distinguished from generalizability of the CNLP/MPSE workflow. The CNLP portion of the pipeline can be used to create a similar dataset for any institution engaged in WGS-based diagnosis, and, because it is a bayesian classifier, retraining MPSE using these data is straightforward. While we chose to use the CLiX CNLP tool, any NLP software able to produce high-fidelity HPO-based phenotype descriptions could be used upstream of MPSE. Going forward, we will explore the utility of retraining and combining models derived from multi-institutional datasets to further improve performance. Recent work has also demonstrated the utility of WGS for pediatric intensive care unit (PICU) patients, where genome-based diagnoses have ended years-long diagnostic odysseys [24]. The PICU generally has a more heterogeneous patient population than the NICU, because it includes patients from less than 12 months through 18 years of age, and a broader array of medical conditions such as cancer, organ transplant, and trauma. Thus, an automated tool such as MPSE that could help identify the relatively less common percentage of PICU patients with underlying Mendelian disorders could be especially useful for this population. These facts suggest that large medical systems may have other, non-pediatric patients who would also benefit from WGS—if they could be found. MPSE could in principle be used to search electronic medical record databases for such patients. Outpatient pediatric specialty clinics might also benefit from using this type of automated tool.

Re-analysis of previously negative WGS cases is also increasingly an issue. The last decade has witnessed a huge increase in numbers of genes and variants associated with Mendelian conditions [25, 26], with 250 newly described disorders annually, suggesting that many individuals previously undiagnosed by gene panels, WES, and WGS, could benefit from reanalysis in light of our ever-expanding knowledge of genetic disease. Recent work has validated this hypothesis [27, 28]. However, limited reimbursement and resources mean that, to be cost-effective, only those patients with the highest likelihood of diagnosis are currently reanalyzed using WGS technologies. Once again, automated approaches such as the one described here might provide a means to locate and prioritize these patients for reanalysis. High MPSE scores might also be used to strengthen arguments for reimbursement. More generally, we foresee MPSE as an electronic decision support tool for facilitating the patient review process.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-023-01166-7.

Additional file 1: Table S1. Clinical diagnosis frequencies for sequenced RCHSD cases broken down by positive/negative molecular diagnosis status. **Table S2.** Primary clinical diagnoses for sequenced Utah NeoSeq cases broken down by positive/negative molecular diagnosis status. **Table S3.** Individual MPSE scores for RCHSD cohort patients. **Table S4.** Individual MPSE scores for Utah NeoSeq cohort patients.

Acknowledgements
We thank Aaron Quinlan and Gabor Marth for many helpful discussions in the early stages of the project.

Authors' contributions

MY, BP, CH, and SK wrote the manuscript. BP, JH MY, CH, SK, SMJ, and LB designed the study and analysis strategy. MY, BP, and JH developed the MPSE algorithm. EF, CH, SMB, and MY guided requirements. MM, BP, JH, SZ, and MY performed the data analysis. EJ, ESK, and MNB compiled cases and clinical evidence. CH, EF, and EJ provided feedback on features and development. AO, SMJ, LB, and SK sponsored the project and provided helpful discussions and edits of the manuscript. The authors read and approved the final manuscript.

Funding
The preparation of this manuscript was supported by a grant from the Conrad Prebys Foundation and by NIH grants UL1TR002550 from NCATS to E.J. Topol (with sub-award to Rady Children's Institute for Genomic Medicine). The Utah NeoSeq Project was funded by the Center for Genomic Medicine at the University of Utah Health, ARUP Laboratories, the Ben B. and Iris M. Margolis Foundation, the R. Harold Burton Foundation, and the Mark Miller Foundation. This work utilized resources and support from the Center for

High Performance Computing at the University of Utah. The computational resources used were partially funded by the NIH Shared Instrumentation grant 1S10OD021644-01A1.

Availability of data and materials

Due to patient privacy, data sharing consent, and HIPAA regulations, the raw data used in this study cannot be submitted to publicly available databases. However, anonymized output from MPSE for all patients reported here are tabulated in Additional file 1: Tables S3 and S4. MPSE source code, documentation, and synthetic datasets are available on GitHub (https://github.com/Yandell-Lab/MPSE) [15]. No new WGS data are presented in this study.

Declarations

Ethics approval and consent to participate

The need for Institutional Review Board Approval at Rady Children's Hospital for the current study was waived as all data used from this project had previously been generated as part of IRB approved studies and none of the results reported in this manuscript can be used to identify individual patients. The studies from which cases were derived were previously approved by the Institutional Review Boards of Rady Children's Hospital. The University of Utah Institutional Review Board approved the use of human subjects for this research, under a waiver for the requirement to obtain informed consent.

Consent for publication

Not applicable. All patient data presented is de-identified.

Competing interests

EF is an employee of Fabric Genomics Inc.. MY is a consultant to Fabric Genomics Inc., which has a co-marketing agreement with Clinithink Inc. BM and JH have received consulting fees and stock grants from Fabric Genomics Inc. The remaining authors declare that they have no competing interests.

Author details

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA. [2]Department of Human Genetics, Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA. [3]Rady Children's Institute for Genomic Medicine, San Diego, CA, USA. [4]Division of Neonatology, Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT, USA. [5]Department of Pediatrics, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [6]Fabric Genomics Inc., Oakland, CA, USA. [7]Rady Children's Hospital, San Diego, CA, USA.

Received: 25 August 2022 Accepted: 27 February 2023
Published online: 16 March 2023

References

1. Church G. Compelling Reasons for repairing human germlines. N Engl J Med. 2017;377(20):1909–11. https://doi.org/10.1056/NEJMp1710370.
2. Farnaes L, Hildreth A, Sweeney NM, et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. NPJ Genomic Med. 2018;3:10. https://doi.org/10.1038/s41525-018-0049-4.
3. Petrikin JE, Cakici JA, Clark MM, et al. The NSIGHT1-randomized controlled trial: rapid whole-genome sequencing for accelerated etiologic diagnosis in critically ill infants. NPJ Genomic Med. 2018;3:6. https://doi.org/10.1038/s41525-018-0045-8.
4. French CE, Delon I, Dolling H, et al. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. Intensive Care Med. 2019;45(5):627–36. https://doi.org/10.1007/s00134-019-05552-x.
5. Kingsmore SF, Cakici JA, Clark MM, et al. A randomized, controlled trial of the analytic and diagnostic performance of singleton and trio, rapid genome and exome sequencing in ill infants. Am J Hum Genet. 2019;105(4):719–33. https://doi.org/10.1016/j.ajhg.2019.08.009.
6. 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):1231–8. https://doi.org/10.1016/j.ajhg.2021.05.008.
7. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Med. 2015;7(1):81. https://doi.org/10.1186/s13073-015-0199-2.
8. Singleton MV, Guthery SL, Voelkerding KV, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014;94(4):599–610. https://doi.org/10.1016/j.ajhg.2014.03.010.
9. Cipriani V, Pontikos N, Arno G, et al. An improved phenotype-driven tool for rare mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. Genes. 2020;11(4). https://doi.org/10.3390/genes11040460.
10. Birgmeier J, Haeussler M, Deisseroth CA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020;12(544):eaau9113. https://doi.org/10.1126/scitranslmed.aau9113.
11. Groza T, Köhler S, Moldenhauer D, et al. The human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet. 2015;97(1):111–24. https://doi.org/10.1016/j.ajhg.2015.05.020.
12. Clark MM, Hildreth A, Batalov S, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med. 2019;11(489):eaat6177. https://doi.org/10.1126/scitranslmed.aat6177.
13. James KN, Clark MM, Camp B, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. NPJ Genomic Med. 2020;5(1):1–8. https://doi.org/10.1038/s41525-020-00140-1.
14. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med. 2021;13(1):153. https://doi.org/10.1186/s13073-021-00965-0.
15. Peterson B, Hernandez J, Hobbs C, et al. Mendelian Phenotype Search Engine 2023. https://github.com/Yandell-Lab/MPSE
16. 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):942–52. https://doi.org/10.1016/j.ajhg.2020.10.003.
17. Sweeney NM, Nahas SA, Chowdhury S, et al. Rapid whole genome sequencing impacts care and resource utilization in infants with congenital heart disease. NPJ Genomic Med. 2021;6(1):29. https://doi.org/10.1038/s41525-021-00192-x.
18. Nicholas TJ, Al-Sweel N, Farrell A, et al. Comprehensive variant calling from whole-genome sequencing identifies a complex inversion that disrupts ZFPM2 in familial congenital diaphragmatic hernia. Mol Genet Genomic Med. 2022;10(4):e1888. https://doi.org/10.1002/mgg3.1888.
19. Clinithink. Clinithink: AI Solutions Company, Clinical Data Solutions for Life Science & Healthcare. Accessed March 5, 2021. https://www.clinithink.com.
20. Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Adv Neural Inf Process Syst. 2001;14:8.
21. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(85):2825–30.
22. Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning. 1st ed. New York: Springer; 2001. https://link.springer.com/book/10.1007/978-0-387-21606-5. Accessed 20 Apr 2022
23. 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. 2019;21(7):1585–93. https://doi.org/10.1038/s41436-018-0381-1.
24. Sanford EF, Clark MM, Farnaes L, et al. Rapid whole genome sequencing has clinical utility in children in the PICU. Pediatr Crit Care Med J Soc Crit Care Med World Fed Pediatr Intensive Crit Care Soc. 2019;20(11):1007–20. https://doi.org/10.1097/PCC.0000000000002056.
25. Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. Am J Hum Genet. 2019;105(3):448–55. https://doi.org/10.1016/j.ajhg.2019.07.011.
26. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789–98. https://doi.org/10.1093/nar/gku1205.

27. Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing data. N Engl J Med. 2019;380(25):2478–80. https://doi.org/10.1056/NEJMc1812033.
28. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med Off J Am Coll Med Genet. 2017;19(2):209–14. https://doi.org/10.1038/gim.2016.88.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.