MPSE: The Mendelian Phenotype Search Engine, a Beginner's Guide.

Bennet Peterson[1], Edwin F. Juarez[2], Edgar Javier Hernandez[1], Charlotte Hobbs[2], Matthew N. Bainbridge[2], Mark Yandell[1]*

[1] Department of Human Genetics, Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA, [2] Rady Children's Institute for Genomic Medicine, San Diego, CA, USA.

* Corresponding author, myandell@genetics.utah.edu

Automatically prioritize patients for WGS. NICU, PICU and specialty clinics provide powerful recruitment points for our Mendelian disease discovery efforts. However, manually searching patient medical histories to identify the best candidates for WGS is a time consuming, cumbersome, and largely *ad hoc* process. Automated means to continuously survey a NICU, or even an entire hospital system for those patients most likely to have undiagnosed Medelian diseases will improve care, and speed discovery (**Fig. 1**). Toward this end, Rady's Children's Hospital, and the U of Utah have begun to explore the possibilities of using Natural Language Processing tools (NLP) to directly convert clinic notes and adjunct EHR data into machine readable, Human Phenotype Ontology (HPO)-based patient phenotype descriptions. HPO-based phenotyping data are highly desirable, as they provide means for prioritizing patients for WGS (as we show below) and can be directly combined with WGS data GEM [1], for genetic diagnosis and discovery purposes **Our tool for automatically prioritizing patients for WGS is called MPSE—the Mendelian Phenotype Search Engine** [2,3].

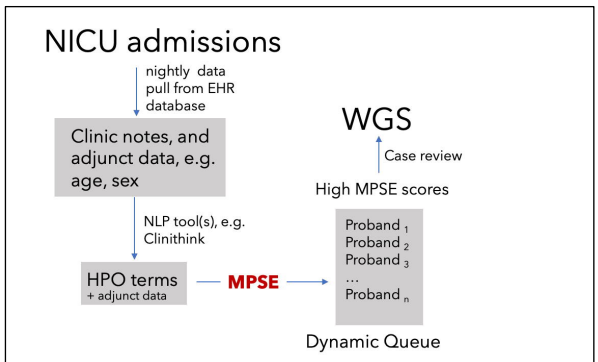

Figure 1 The MPSE workflow. MPSE provides automated means for continuous surveillance, identification, and prioritization of patients with likely Mendelian diseases for Whole Genome Sequencing (WGS).

Benchmark dataset. To investigate the feasibility of this sub aim**,** we took advantage of a unique dataset made available from RCIGM, consisting of 1075 Level IV NICU admits, their clinic notes, and metadata such as age, and gender. 294 of these children had been selected by RADY's for WGS; and 84 were diagnosed with Mendelian diseases. Manual, physician created HPO descriptions are also available for the 294 children with WGS, making this dataset an ideal for proof of principle analyses See [2] for additional details. **Validation Dataset.** Our validation dataset is composed of 2965 University of Utah Level-III NICU admits, and 35 WGS probands sequenced by the University of Utah NeoSeq program [5].

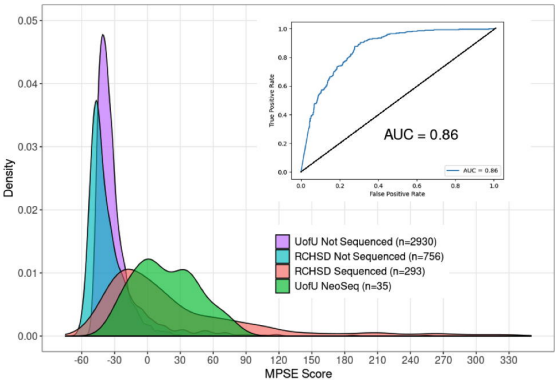

Figure 2. Automatically identifying probands with Mendelian phenotypes and prioritizing them for WGS using NLP-derived HPO phenotype descriptions. Panel A: distributions of MPSE raw scores for RCHSD sequenced (red), and RCHSD unsequenced (blue) probands. Score distributions for Utah NeoSeq (green) and Utah unsequenced probands (purple). Insert: Receiver Operator Characteristic (ROC) curve for RCHSD data. MPSE Scores are -log likelihood ratios.

Prioritization Benchmarks. As can be seen in **Figure 2** , MPSE is both transportable, and effective—when trained the clinical notes of sequenced RCIGM probands, MPSE assigns low prioritization scores to Utah level 3 NICU admits (True Negatives), but high scores to those Utah probands selected for sequencing (green) sequencing. See [1] for additional details. The insert shows a Receiver Operator Characteristic (ROC) curve for the RCIGM 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 RCIGM result (ROC curve not shown), indicating that MPSE provides effective means to prioritize probands for WGS.

MPSE can identify Mendelian Disease within hours of NICU admission. Our initial work [2] demonstrated MPSE's ability to accurately identify sequencing candidates by aggregating clinical notes across from the entirety of the patient's NICU stay up until the time physicians selected them for sequencing. Our second prospective study [3] has shown that an MPSE-based classifier can automatically flag patients for sequencing, resulting in earlier sequencing and faster diagnosis, without decreasing diagnostic yields. Obviously, means to identify WGS candidates as soon as possible upon NICU admission would significantly enhance care, enabling earlier diagnosis mendelian disease and more timely interventions. To examine this possibility, we calculated MPSE scores daily for each patient in our U of Utah cohort using only HPO terms extracted from clinical notes present in the EHR at the given moment. Thus, each patient had a series of MPSE scores for each day spent in the NICU from admission to discharge. Longitudinal MPSE scores for diagnostic, non-diagnostic, and unsequenced patients are shown in **Figure 3**. By the end of the first day (admission day 0) in the NICU, sequenced cases already had significantly higher MPSE scores than unsequenced controls (control mean: -48.4; case mean: -30.6; p=5.4e-10). This trend was consistent and statistically significant throughout the first 30 days post-admission (data not shown). Additionally, sequenced cases saw greater average daily increases in MPSE score than unsequenced controls throughout the first 30 days post-admission. Importantly, the greatest differences between cases and controls in MPSE scores occurs during the first day post-admission, with average case MPSE score rising by 11.9 points and average control MPSE score rising by only 2.7 points (p=4.0e-5).

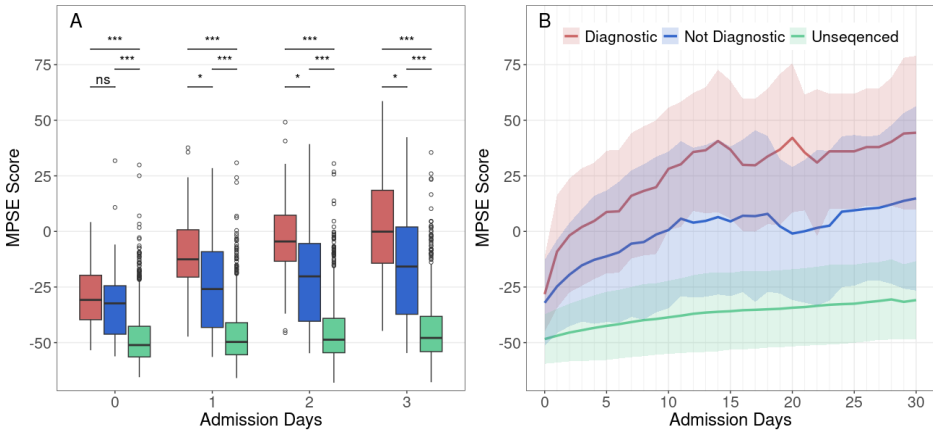

Figure 3. MPSE can identify Mendelian disease within the first 24 hours of NICU admission. **A.** Diagnostic and non-diagnostic sequenced patients had significantly higher MPSE scores than unsequenced patients beginning on admission day 0. Diagnostic patients had significantly higher MPSE scores than non-diagnostic patients beginning on admission day 1, with both trends further strengthening thereafter. Boxplot comparison significance levels: *** (p < 1e-5); * (p < 0.05). Diagnostic (red) and non-diagnostic (blue) NeoSeq patients compared to unsequenced control patients (green). **B.** Trajectory of MPSE scores for diagnostic (red) and non-diagnostic (blue) NeoSeq patients compared to unsequenced control patients (green) during the first 30 days after NICU admission. Solid lines show the mean MPSE score per group and the shaded regions cover one standard deviation from each mean.

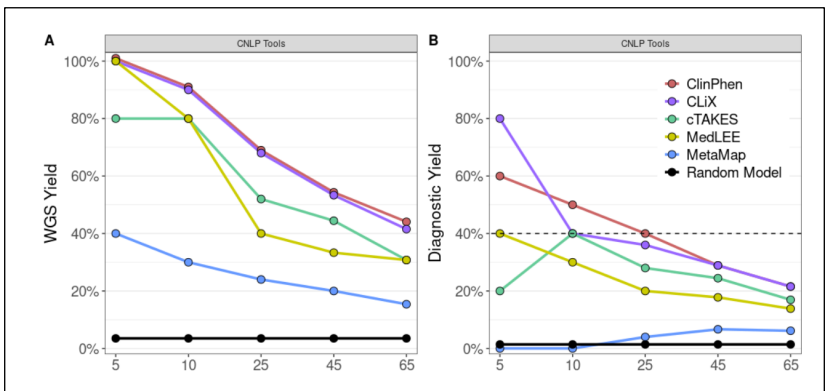

Fig. 4. MPSE works well, regardless of NLP tool. MPSE revalidated using HPO terms provided by different NLP tools. Note that performance using the Open-source tool ClinPhen compares favorably to that of CLiX. Panels A and B display the recovery rates for the 65 sequenced cases. Left. Whole genome sequencing (WGS) yield; Right. Diagnostic yield. Data are for or 1,838 U of Utah NICU patients and five different CNLP tools. Dotted line in right panel indicates the NeoSeq study's total diagnostic yield of 43%. Random Model denotes rate produced by choosing randomly.

<u>Collectively, these results demonstrate that MPSE can effectively prioritize probands for WGS within hours of admission to the NICU; and that probands with the highest MPSE scores also tend to have diagnosable Mendelian conditions</u>.

Implementing MPSE in your Health System. The published benchmarks [2,3] of MPSE employ a Clinithink CLiX™ based NLP pipeline. CLiX™ is a commercial tool.

Recognizing that licensing a commercial tool can be expensive and time consuming, we recently carried out additional benchmarks using a variety of open-source tools (**Fig 4**) [4]. As can be seen, performance is good using a variety of these NLP tools, with that obtained using ClinPen rivaling CLiX. We have also measured performance using different subsets of notes and investigated the impact of duplicated notes on MPSE performance, reporting that the impact of these factors is minimal [2, 3], meaning that simply grabbing every available note works well.

ClinPhen and MPSE are entirely open-source projects, thus barriers to deploy these tools in your health system is minimal. All that is required is means to retrieve clinic notes from the institution's EHR database and pass the notes—even redundant notes to the NLP tool of choice. MPSE will do the rest.

MPSE is now powering the Gene Kids project, a new clinical-research initiative of the Primary Children's Hospital Center for Personalized Medicine and the University of Utah made possible by a $9 million grant from the Warren Alpert foundation and Intermountain Heath Care. Primary Children's Hospital cares for over 1.7 million children across the Intermountain West, the largest pediatric catchment area in the nation [7]. If MPSE can do this, it can power your project too.

MPSE and related publications.

1. De La Vega FM, Chowdhury S, Moore B, Frise E, McCarthy J, Hernandez EJ, Wong T, James K, Guidugli L, Agrawal PB, Genetti CA, Brownstein CA, Beggs AH, Löscher BS, Franke A, Boone B, Levy SE, Õunap K, Pajusalu S, Huentelman M, Ramsey K, Naymik M, Narayanan V, Veeraraghavan N, Billings P, Reese MG, Yandell M, Kingsmore SF. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med. 2021 Oct 14;13(1):153. doi: 10.1186/s13073-021-00965-0. PMID: 34645491; PMCID: PMC8515723.
2. Peterson B, Hernandez EJ, Hobbs C, Malone Jenkins S, Moore B, Rosales E, Zoucha S, Sanford E, Bainbridge MN, Frise E, Oriol A, Brunelli L, Kingsmore SF, Yandell M. Automated prioritization of sick newborns for whole genome sequencing using clinical natural language processing and machine learning. Genome Med. 2023 Mar 16;15(1):18. doi: 10.1186/s13073-023-01166-7. PMID: 36927505; PMCID: PMC10018992.
3. A Machine Learning Decision Support Tool Optimizes Whole Genome Sequencing Utilization in a Neonatal Intensive Care Unit. Edwin F. Juarez, Bennet Peterson, Erica Sanford Kobayashi, Sheldon Gilmer, Laura E. Tobin, Brandan Schultz, Jerica Lenberg, Jeanne Carroll, Shiyu Bai-Tong, Nathaly M. Sweeney, Curtis Beebe, Lawrence Stewart, Lauren Olsen, Julie Reinke, Elizabeth A. Kiernan, Rebecca Reimers, Kristen Wigby, Chris Tackaberry, Mark Yandell, Charlotte Hobbs, Matthew N. Bainbridge. medRxiv 2024.07.05.24310008; doi: https://doi.org/10.1101/2024.07.05.24310008
4. MPSE: Fast and flexible prioritization of critically ill newborns for whole genome sequencing. Peterson B, Hernandez EJ, Bainbridge, MN, Yandell, M. Manuscript in preparation.
5. NeoSeq program. https://uofuhealth.utah.edu/center-genomic-medicine/research/utah-neoseq-project
6. U of U Gene Kids press release. https://uofuhealth.utah.edu/newsroom/news/2024/08/new-collaboration-aims-provide-genetic-diagnoses-thousands-of-kids