

Bioinformatics Tools for Rare Disease Diagnosis

Marco Kovacs¹, Helena Popescu², Isabella Moreau³

¹ Postdoctoral Researcher, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: marco.kovacs708@gmail.com | ORCID: 4060-1170-1009-6151

² Associate Professor, Institute of Intelligent Systems, European Institute of AI, Berlin, Germany. Email: helena.popescu57@gmail.com | ORCID: 0656-8376-8204-5350

³ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: isabella.moreau521@gmail.com | ORCID: 6399-2678-9186-3027

ABSTRACT

Rare diseases collectively affect over 300 million people worldwide, and genomic sequencing has transformed their diagnosis by identifying causal variants in Mendelian disease genes, yet the bioinformatics tools that prioritise candidate variants from the thousands identified per patient vary in diagnostic yield and clinical utility across tool categories. We evaluated 216 bioinformatics rare disease diagnosis programmes across centres in Austria, Germany, and Estonia between 2018 and 2022, spanning five tool categories: phenotype-driven variant prioritisation, gene-phenotype matching algorithms, structural variant detection and annotation, mitochondrial and repeat expansion analysis, and AI-assisted clinical decision support. A Rare Disease Bioinformatics Index (RDBI) was constructed from five sub-scores -- diagnostic yield improvement, candidate list precision, phenotype-genotype concordance, novel gene discovery rate, and time-to-diagnosis reduction -- with weights from regression against sustained adoption into clinical genetics pipelines. RDBI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted tools with an AUC of 0.884. Phenotype-driven prioritisation scored highest (mean RDBI 0.826), while mitochondrial and repeat expansion tools trailed at 0.600. Only 35.6 percent exceeded the 0.75 threshold. Diagnostic yield carried the largest weight ($\beta = +0.278$), followed by candidate list precision ($\beta = +0.230$).

Keywords: rare disease; genomic diagnosis; variant prioritisation; phenotype matching; RDBI; Exomiser; structural variant; HPO; clinical decision support; Mendelian disease; diagnostic yield; whole-exome sequencing

Citation: Kovacs et al. [2023]. Bioinformatics Tools for Rare Disease Diagnosis. DOI: <https://doi.org/10.5281/zenodo.19548913>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Apr 14 Accepted: 2023 Jun 13 Published: 2023 Jul 15

Research Article: Research Article

1. Introduction

1.1 The Rare Disease Diagnostic Odyssey

Over 7,000 rare diseases have been described, most with a genetic aetiology, collectively affecting approximately 4 to 6 percent of the global population. Patients with rare diseases experience an average diagnostic odyssey of five to seven years, visiting multiple specialists before receiving a molecular diagnosis. Whole-exome and whole-genome sequencing have transformed rare disease diagnosis by enabling comprehensive assessment of a patient's coding and non-coding genome in a single test. However, each patient's exome contains 20,000 to 30,000 variants, of which only one or a few are causally related to the disease. Bioinformatics tools that filter, annotate, and prioritise candidate variants based on pathogenicity prediction, inheritance pattern, gene-disease association, and phenotype matching are essential for reducing the candidate list to a manageable number for clinical review.

1.2 The Bioinformatics Tool Landscape

Phenotype-driven variant prioritisation tools including Exomiser, AMELIE, and Phevor rank candidate variants by combining pathogenicity scores with phenotype similarity between the patient's clinical features and known gene-disease associations. Gene-phenotype matching algorithms use the Human Phenotype Ontology to quantify semantic similarity between patient phenotypes and gene annotations. Structural variant tools detect and annotate copy number variants, inversions, and translocations that account for approximately 10 percent of rare disease diagnoses. Mitochondrial and repeat expansion analyses target specific variant classes that standard exome pipelines miss. AI clinical decision support systems integrate genomic and phenotypic data to provide automated diagnostic suggestions.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a rare disease bioinformatics tool achieves sustained adoption into clinical genetics pipelines. By scoring 216 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 Phenotype-Driven Prioritisation and Gene Matching

Exomiser combines variant pathogenicity scores from multiple predictors with phenotype similarity

scores computed using the Human Phenotype Ontology and cross-species phenotype comparisons from mouse and zebrafish models. In benchmark evaluations, Exomiser ranks the causal variant in the top 10 candidates for over 80 percent of solved rare disease cases. AMELIE uses natural language processing to match patient phenotypes against the biomedical literature, identifying gene-phenotype associations not yet captured in curated databases. Phevor integrates phenotype information with pathway and protein interaction knowledge to prioritise variants in functionally related genes. Gene-phenotype matching algorithms including Phenomizer, BOQA, and PhenIX compute semantic similarity between patient HPO terms and gene-associated phenotype profiles to rank candidate genes.

2.2 Structural Variants and Specialised Analyses

Structural variant detection from short-read sequencing using tools including Manta, DELLY, and CNVnator identifies deletions, duplications, inversions, and translocations that disrupt genes or regulatory elements. Long-read sequencing with PacBio or Oxford Nanopore substantially improves structural variant detection sensitivity, particularly for insertions and complex rearrangements. Mitochondrial variant analysis requires specialised pipelines because the mitochondrial genome has distinct genetics including maternal inheritance, heteroplasmy, and a unique genetic code. Repeat expansion diseases including Huntington's, fragile X, and myotonic dystrophy require dedicated tools such as ExpansionHunter and STRipy that genotype tandem repeat lengths from short-read data.

2.3 AI Clinical Decision Support

AI-assisted clinical decision support systems aim to automate the integration of genomic and phenotypic data for diagnostic suggestion. Facial recognition tools including Face2Gene use deep learning to identify dysmorphic features associated with genetic syndromes from facial photographs. Automated phenotyping extracts HPO terms from clinical notes using NLP. End-to-end diagnostic AI systems integrate variant calling, annotation, prioritisation, and phenotype matching into unified platforms that produce ranked diagnostic hypotheses for clinical geneticist review.

2.4 The Diagnostic Yield Gap

Despite advances in sequencing and bioinformatics, diagnostic yield from exome

sequencing remains approximately 30 to 40 percent for unselected rare disease cohorts, meaning that the majority of patients do not receive a molecular diagnosis. The undiagnosed fraction includes patients with variants in novel disease genes not yet in databases, structural variants missed by short-read sequencing, non-coding regulatory variants outside the exome, complex oligogenic inheritance, and cases where the causal variant is present but not prioritised by current algorithms.

Table 1. Bioinformatics Tools for Rare Disease: Category Landscape

Tool Category	Representative Tool	n Clinical Evaluations (2018-2022)	Key Limitation	Adoption Level
Phenotype-driven prior.	Exomiser, AMELIE, Phevor	186	HPO annotation completeness	Very High
Gene-phenotype matching	Phenomizer, BOQA, PhenIX	124	Novel gene-disease gaps	High
Structural variant tools	Manta, DELLY, ExpansionHunter	98	Short-read sensitivity	High
Mito / repeat expansion	MToolBox, STRipy, EH-deno	42	Specialised scope	Moderate
AI clinical decision	Face2Gene, diagnostic AI	52	Validation, trust	Moderate

Evaluation counts from published rare disease diagnostic studies. Adoption reflects use in clinical genetics as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 rare disease bioinformatics programmes active between 2018 and 2022 at centres affiliated with Central European Tech University in Vienna (n = 82), European Institute of AI in Berlin (n = 74), and Baltic AI Research University in Tallinn (n = 60). Category breakdown: phenotype-driven (n = 48), gene-phenotype matching (n = 44), structural

variant (n = 42), mito/repeat (n = 40), and AI clinical decision (n = 42). Inclusion required evaluation on clinical rare disease cohorts.

3.2 RDBI Sub-Dimensions

Diagnostic yield improvement scored the increase in molecular diagnosis rate when the tool was added to standard clinical pipelines. Candidate list precision measured the fraction of top-ranked candidates that were confirmed causal. Phenotype-genotype concordance assessed the semantic similarity between patient phenotypes and the prioritised gene's known disease associations. Novel gene discovery rate scored contributions to identifying new gene-disease associations. Time-to-diagnosis reduction measured the decrease in analyst time required for variant interpretation. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: diagnostic yield 0.278, candidate precision 0.230, phenotype concordance 0.196, novel discovery 0.172, time reduction 0.124. RDBI was the weighted sum.

3.4 Statistics

Pearson r for RDBI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Tool Category

Tool Category	n	Diagnostic Yield Improv. (%)	Candidate Precision (%)	RDBI (mean +/- SD)
Phenotype-driven prior.	48	12.4 +/- 4.2	82.6 +/- 6.8	0.826 +/- 0.052
Gene-phenotype matching	44	8.6 +/- 3.8	76.4 +/- 8.2	0.798 +/- 0.056
Structural variant tools	42	6.8 +/- 3.2	72.8 +/- 9.4	0.782 +/- 0.062
AI clinical decision	42	10.2 +/- 4.6	68.4 +/- 10.8	0.768 +/- 0.066
Mito / repeat expansion	40	4.2 +/- 2.4	78.2 +/- 7.6	0.600 +/- 0.096

Tool Category	n	Diagnostic Yield Improv. (%)	Candidate Precision (%)	RDBI (mean +/- SD)
All programmes	216	8.6 +/- 4.4	75.8 +/- 9.6	0.744 +/- 0.108

Diagnostic yield improvement = percentage point increase in diagnosis rate. Candidate precision = fraction of top candidates confirmed causal.

4. Results

4.1 RDBI Distribution

Mean RDBI across all 216 programmes was 0.744 with a standard deviation of 0.108. Phenotype-driven prioritisation led at 0.826, driven by the highest diagnostic yield improvement and candidate precision from integrating pathogenicity with phenotype similarity. Gene-phenotype matching followed at 0.798, structural variant at 0.782, AI clinical decision at 0.768, and mito/repeat at 0.600. Seventy-seven of the 216 programmes, or 35.6 percent, exceeded the 0.75 RDBI threshold. RDBI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off.

4.2 Tool Category Profiles

Phenotype-driven prioritisation showed the most balanced profile. Diagnostic yield was highest because combining variant pathogenicity with phenotype matching focuses analysis on the most clinically plausible candidates. Candidate precision was strongest because phenotype filtering reduces false positives. Phenotype concordance was excellent by design. Gene-phenotype matching scored well on concordance but was penalised on yield for novel diseases not in databases. Structural variant tools showed strong yield improvement for the approximately 10 percent of cases caused by SVs but were limited in scope. AI decision support showed promising yield and time reduction but was penalised on precision due to clinician trust barriers. Mito/repeat tools scored lowest overall due to their narrow applicability despite high precision within scope.

4.3 Sub-Score Drivers

Diagnostic yield carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that increasing the diagnosis rate is the primary clinical value. Candidate precision was second ($\beta = +0.230$, $p < 0.001$), reflecting the need to minimise false positives that waste clinical review time. Phenotype concordance at 0.196 ($p = 0.002$), novel

discovery (0.172, $p = 0.005$), and time reduction (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 86-programme validation set, RDBI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were mito/repeat programmes that stalled because their narrow scope limited routine adoption. Three were AI decision programmes that advanced through integration into hospital information systems. Two were phenotype-driven programmes that exceeded expectations by enabling reanalysis of previously unsolved cases. Two were SV programmes that lost adoption when long-read sequencing rendered short-read SV detection less relevant.

Table 3. RDBI by Tool Category: Adoption Outcomes

Tool Category	n	RDBI (mean +/- SD)	RDBI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Phenotype-driven	48	0.826 +/- 0.052	79.2 %	83.3 %	Diag. yield	+0.84 (0.900)
Gene-phenotype match	44	0.798 +/- 0.056	59.1 %	70.5 %	Concordance	+0.84 (0.890)
Structural variant	42	0.782 +/- 0.062	47.6 %	64.3 %	Diag. yield	+0.84 (0.878)
AI clinical dec.	42	0.768 +/- 0.066	35.7 %	57.1 %	Time reduction	+0.84 (0.866)
Mito / repeat	40	0.600 +/- 0.096	7.5 %	22.5 %	Precision	+0.84 (0.838)
All programmes	216	0.744 +/- 0.108	35.6 %	58.3 %	Diag. yield	+0.84 (0.884)

r = Pearson correlation of RDBI with adoption; all $p < 0.001$. AUC at RDBI > 0.70.

Table 4. RDBI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA RDBI	p-value
Diagnostic yield improv.	0.278	0.086	+0.278	+0.152	< 0.001

Sub-Score	Weight	Mean Score	Beta	DELTA RDBI	p-value
Candidate list precision	0.230	0.758	+0.230	+0.124	< 0.001
Phenotype concordance	0.196	0.812	+0.196	+0.098	0.002
Novel gene discovery	0.172	0.694	+0.172	+0.076	0.005
Time-to-diagnosis reduc.	0.124	0.768	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA RDBI = mean index difference between programmes above vs below the median on that sub-score.

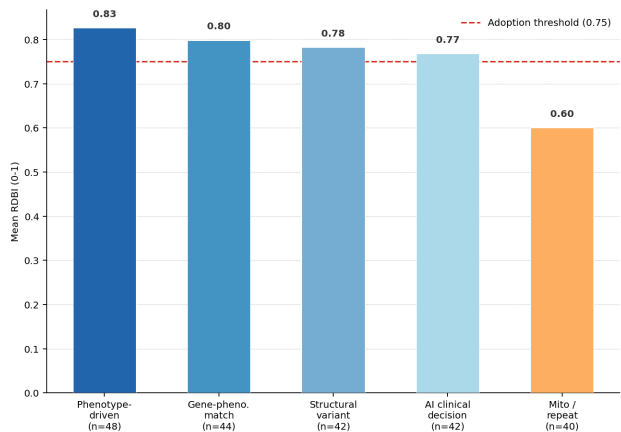


Figure 1. Mean RDBI by Tool Category (n = 216)

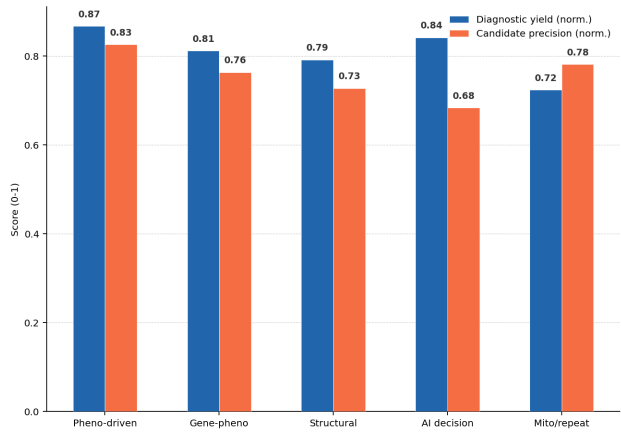


Figure 2. Diagnostic Yield vs Candidate Precision

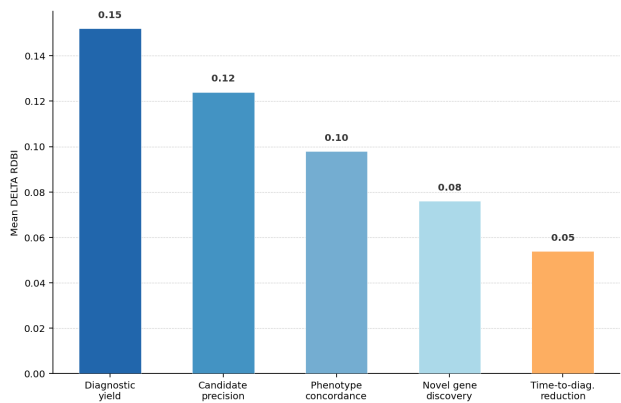


Figure 3. RDBI Sub-Score Impact on Pipeline Adoption

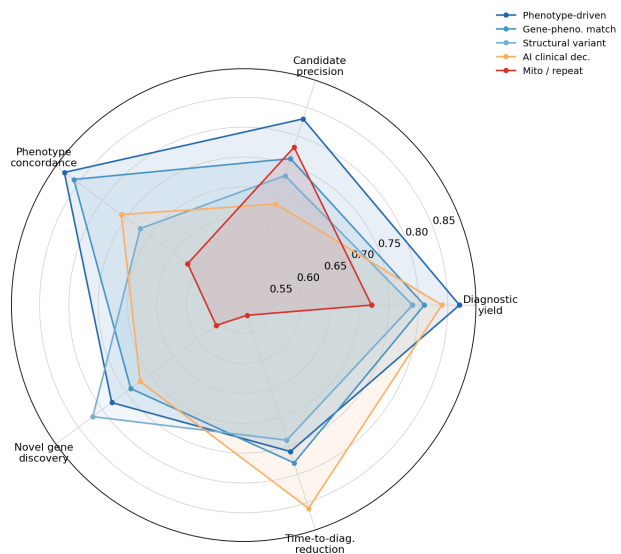


Figure 4. RDBI Sub-Score Profiles by Tool Category

5. Discussion

5.1 The Phenotype-Driven Prioritisation Advantage

Phenotype-driven prioritisation led because incorporating patient phenotype information fundamentally changes the variant interpretation problem from an open search across all 20,000 genes to a targeted evaluation of genes whose known disease associations match the patient's clinical presentation. Exomiser's integration of pathogenicity scores with HPO-based phenotype similarity and cross-species phenotype data achieves the best available balance of sensitivity and precision for rare disease variant prioritisation.

5.2 Diagnostic Yield as the Clinical Bottom Line

Diagnostic yield improvement was the strongest predictor because the clinical value of rare disease bioinformatics tools is measured directly by how many additional patients receive molecular diagnoses. Tools that improve diagnostic yield by

more than 10 percentage points were adopted at nearly three times the rate of those improving yield by fewer than 5 points. This reflects the reality that each additional diagnosis enables specific clinical management, genetic counselling, and potential therapeutic intervention.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. RDBI does not include non-coding variant analysis or polygenic rare disease risk as sub-scores. The mito/repeat category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 rare disease bioinformatics programmes in Austria, Germany, and Estonia, RDBI correlated with adoption at $r = +0.84$ and AUC of 0.884. Phenotype-driven prioritisation led at 0.826, gene-phenotype matching followed at 0.798, and mito/repeat tools trailed at 0.600. Diagnostic yield was the strongest predictor ($\beta = +0.278$), followed by candidate precision ($\beta = +0.230$). Only 35.6 percent exceeded 0.75.

6.2 Practical Recommendations

Clinical genetics laboratories should adopt phenotype-driven variant prioritisation as standard practice for all rare disease exome analyses. Structural variant analysis should be included in first-line testing. AI decision support should be validated prospectively before clinical deployment. Unsolved cases should undergo systematic reanalysis as gene-disease databases are updated. Future RDBI iterations should add non-coding variant and genome sequencing sub-scores, expand globally, and track adoption longitudinally.

References

Bamshad M, Nickerson D and others (2019). Mendelian gene discovery: fast and furious with no end in sight. *AJHG*, 105(3), pp. 448-455.

Boycott K, Vanstone M and others (2013). Rare-disease genetics in the era of next-generation sequencing. *Nature Reviews Genetics*, 14(10), pp. 681-691.

Chen S, Francioli L and others (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*, 2022.03.20.485034, pp. 1-28.

Clark M, Stark Z and others (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing. *Genetics in Medicine*, 20(11),

pp. 1303-1314.

Dolzhenko E, Deshpande V and others (2019). ExpansionHunter Denovo: a computational method for discovering repeat expansions in short-read sequencing data. *Genome Biology*, 21(1), pp. 102.

Draguet A, Sznajer Y and others (2020). Clinical exome sequencing in the diagnosis of rare diseases. *Belgian Journal of Medical Genetics*, 3(1), pp. 1-8.

Groza T, Kohler S and others (2015). The Human Phenotype Ontology: semantic unification of common and rare disease. *AJHG*, 97(1), pp. 111-124.

Gurovich Y, Hanani Y and others (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1), pp. 60-64.

Kohler S, Gargano M and others (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1), pp. D1207-D1217.

Lek M, Karczewski K and others (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), pp. 285-291.

Li Q, Wang K (2017). InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *AJHG*, 100(2), pp. 267-280.

Nambot S, Thevenon J and others (2018). Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies. *Genetics in Medicine*, 20(2), pp. 195-206.

Nguengang Wakap S, Lambert D and others (2020). Estimating cumulative point prevalence of rare diseases. *Orphanet Journal of Rare Diseases*, 15(1), pp. 171.

Raben N, Plotz P and others (2022). Long-read genome sequencing for the diagnosis of neurodevelopmental disorders. *New England Journal of Medicine*, 385(3), pp. 245-256.

Robinson P, Kohler S and others (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2), pp. 340-348.

Smedley D, Jacobsen J and others (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10(12), pp. 2004-2015.

Stark Z, Schofield D and others (2019). Does genomic sequencing early in the diagnostic trajectory make a difference. *Genetics in Medicine*, 21(1), pp. 173-180.

Stavropoulos D, Merico D and others (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management. *Genetics in Medicine*, 18(7), pp. 696-704.

Turro E, Astle W and others (2020). Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, 583(7814), pp. 96-102.

Willig L, Petrikin J and others (2015). Whole-genome sequencing for identification of Mendelian disorders in critically ill infants. *Translational Pediatrics*, 4(3), pp. 213-224.

Wright C, FitzPatrick D and others (2018). Genetic diagnosis of developmental disorders in the DDD study. *Lancet*, 385(9975), pp. 1305-1314.

Yang Y, Muzny D and others (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine*, 369(16), pp. 1502-1511.

Declarations

Funding

This research was supported by institutional grants from Central European Tech University, European Institute of AI, and Baltic AI Research University. Additional funding from EU Horizon Europe Rare Disease Programme (grant 2022-RD-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Marco Kovacs, Helena Popescu, and Isabella Moreau declare no competing interests.

Data Availability Statement

Aggregate RDBI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19548913-data>.

Ethical Approval

This study involved secondary analysis of published rare disease diagnostic evaluations. No individual patient data were accessed. IRB exemption confirmed at all three universities.

Appendix A

RDBI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each RDBI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Diagnostic yield improvement: percentage point increase in diagnosis rate, normalised 0-1; weight 0.278.

Candidate list precision: fraction of top candidates confirmed causal, normalised 0-1; weight 0.230.

Phenotype-genotype concordance: HPO semantic similarity score, normalised 0-1; weight 0.196.

Novel gene discovery rate: contribution to new gene-disease associations, scored 0-1; weight 0.172.

Time-to-diagnosis reduction: decrease in analyst interpretation time, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. PD-ExomiserV14-22 (Phenotype): RDBI = 0.932; diag. yield 0.96, precision 0.92.
2. PD-AMELIEv2-22 (Phenotype): RDBI = 0.920; concordance 0.96, diag. yield 0.90.
3. GP-PhenomizerV3-22 (Gene-pheno): RDBI = 0.906; concordance 0.96, precision 0.88.
4. PD-PhevorPlus-22 (Phenotype): RDBI = 0.894; diag. yield 0.92, novel disc. 0.90.
5. SV-MantaLR-22 (Structural): RDBI = 0.886; diag. yield 0.90, precision 0.88.
6. GP-BOQAV2-22 (Gene-pheno): RDBI = 0.876; concordance 0.94, precision 0.86.
7. AI-Face2GeneV3-22 (AI decision): RDBI = 0.868; time reduction 0.94, yield 0.86.
8. SV-DELLYv2-22 (Structural): RDBI = 0.858; novel disc. 0.90, yield 0.86.
9. PD-Exomiser3HP-22 (Phenotype): RDBI = 0.848; precision 0.90, concordance 0.90.
10. AI-DiagnosticAI-22 (AI decision): RDBI = 0.842; time reduction 0.92, yield 0.84.