WISELEAF SCIENTIFIC VENTURES

SMART HEALTHCARE REVOLUTION: AI STRATEGIES FOR SUPERIOR GLOBAL HEALTH

CHAPTER 7- ARTIFICIAL INTELLIGENCE IN RARE DISEASE DIAGNOSIS

Mr. Lakhan Ravi Bukkawar

Lecturer, Department of Pharmaceutical Science, Ashoka Institute of Pharmacy, Near Synergy World, Kosara, Chandrapur, Maharashtra, India

Email: laakhanbukkawar@gmail.com

ABSTRACT

Rare diseases are individually uncommon but collectively frequent. The Orphanet database lists more than 6,000 of them, about seven in ten are genetic, and together they affect an estimated 3.5 to 5.9% of the world's population at any time. The defining clinical problem is delay: European patients wait on average 4.7 years for a diagnosis, and many never receive one. This chapter examines how artificial intelligence (AI) is being used to shorten that wait. It begins with the computable representation of clinical features through the Human Phenotype Ontology and related resources, because nearly every downstream algorithm depends on structured phenotypes. It then reviews deep-learning facial phenotyping (DeepGestalt, as deployed in Face2Gene, and GestaltMatcher), including evidence that performance varies with ancestry and that systems trained mainly on European faces may serve Indian, African and East Asian patients less well. The central sections deal with genomic diagnosis: pooled diagnostic yields of exome and genome sequencing, the ACMG/AMP variant classification framework, phenotype-driven prioritisation with Exomiser in the 100,000 Genomes Project, and machine-learned variant-effect predictors such as CADD, SpliceAI and AlphaMissense, together with the persistent problem of variants of uncertain significance. Rapid genome sequencing in neonatal intensive care, where automated pipelines have brought provisional diagnoses under a day, is treated separately. Emerging uses of electronic health records, large language models and newborn screening are discussed with attention to the thinness of current evidence, as is knowledge-graph drug repurposing with TxGNN. The Indian context is traced through the National Policy for Rare Diseases 2021 and its Centres of Excellence. The chapter closes with the structural constraints that limit all of this work: tiny datasets, the need for few-shot methods, federated data sharing through Matchmaker Exchange, and equity of access.

Keywords: rare diseases; Human Phenotype Ontology; facial phenotyping; variant prioritisation; rapid genome sequencing; National Policy for Rare Diseases

1. Introduction

A physician in general practice may see a given rare disease once in a career, or never. That single fact explains much of what is difficult about rare disease medicine. Pattern recognition, the core skill of clinical diagnosis, depends on having seen the pattern before. For conditions with a prevalence of one in a million, almost no clinician has. The patient therefore moves from specialist to specialist, accumulating investigations, partial explanations and occasionally wrong labels, a sequence now widely called the diagnostic odyssey.

Two developments over the past fifteen years have changed what is possible. The first is the falling cost of DNA sequencing, which made it practical to examine all protein-coding genes (the exome) or the whole genome of a patient in a single test. The second is the maturing of machine learning, especially deep learning, which can extract useful signal from images, text and sequence at a scale no human analyst could manage. The two are complementary. Sequencing produces too much information: millions of variants per genome, of which perhaps one or two cause the disease. Machine learning supplies ways to rank that information, and to connect it with the patient's clinical features, so that a human expert reviews a short list rather than an unmanageable one. Figure 7.1 shows how these components fit together in a typical diagnostic workflow.

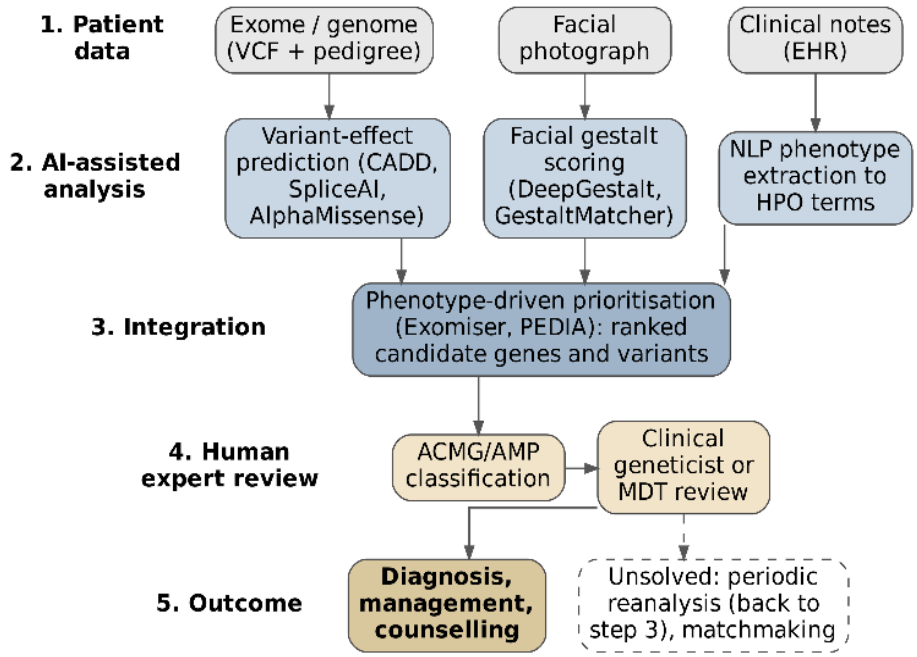

Figure 7.1: An AI-assisted rare disease diagnostic workflow. Phenotypic and genomic inputs are processed by separate AI components, integrated by phenotype-driven prioritisation, and reviewed by human experts; unsolved cases return for periodic reanalysis and matchmaking. Synthesised from [1-4].

The chapter is organised around that division of labour. Section 2 sets out the burden of rare disease and the Indian policy context. Section 3 deals with how phenotypes are made computable, which is the quiet prerequisite for almost everything else. Section 4 covers facial image analysis. Sections 5 and 6 turn to genomic diagnosis, first in routine practice and then in the special setting of the neonatal intensive care unit. Section 7 considers newer and less proven approaches: mining electronic health records, large language models and screening. Section 8 looks briefly at AI-guided drug repurposing, since a diagnosis without a treatment option is only part of the answer. Section 9 discusses the problems that cut across all of these, and Section 10 draws the threads together.

A word on scope. Federated learning as a technical discipline is treated in Chapter 8 of this volume, and the general problems of deploying AI in low-resource health systems in Chapter 10. They are mentioned here only where rare disease raises a distinct issue.

2. The Burden of Rare Disease and the Diagnostic Odyssey

2.1 How many diseases, how many people

Definitions of rarity differ by jurisdiction. The European Union uses a threshold of no more than 5 per 10,000 people, and it was against this threshold that Nguengang Wakap and colleagues analysed the Orphanet database in 2019 [5]. They counted 6,172 unique rare diseases, of which 71.9% were genetic and 69.9% had exclusively paediatric onset. Combining point-prevalence data, they estimated that rare diseases affect 3.5 to 5.9% of the population at any time, or roughly 263 to 446 million people worldwide [5]. The figure is conservative, because it rests only on diseases for which prevalence data exist.

The distribution is extremely skewed. Of the diseases with a recorded point prevalence, 84.5% occur in fewer than one person per million, yet a small group of 149 relatively common rare diseases (4.2% of the total) accounts for about 77 to 81% of the affected population [5]. This matters for AI. The diseases that contribute most patients, and therefore most training data, are not the ones that pose the hardest diagnostic problems. The long tail of ultra-rare conditions is where clinicians most need help and where data-hungry algorithms are least able to learn.

2.2 Why diagnosis takes so long

The best recent European data come from the Rare Barometer survey run by EURORDIS. Among 6,507 respondents living with 1,675 different rare diseases in 41 European countries, the mean total diagnosis time was 4.7 years [6]. Misdiagnosis and the number of professionals consulted were among the factors associated with longer delays [6]. These are respondents from comparatively well-resourced health systems with established genetic services. There is little reason to expect shorter delays where such services are scarce.

The International Rare Diseases Research Consortium (IRDiRC) has set as one of its goals for 2017 to 2027 that any patient with a rare disease already described in the medical literature should be diagnosed within one year of presentation, and that those who cannot be diagnosed should enter a coordinated global research pipeline [7]. With most of that decade gone, the goal has not been met at population scale. It remains a useful yardstick, because it names the two separate tasks that AI must address. One is recognition of a known disease in a new patient. The other is discovery of a disease or gene that has not yet been described.

The cost of delay is concrete. In the 100,000 Genomes Project pilot, a quarter of the genetic diagnoses eventually made had immediate consequences for clinical decision-making [8]. A diagnosis can open access to a specific therapy, end a cascade of futile investigations, inform reproductive choices for the family and connect the family with others who share the same condition.

2.3 The Indian policy context

India has no reliable national prevalence estimate for rare diseases, and the policy discussion has proceeded largely without one. The National Policy for Rare Diseases (NPRD) was launched in March 2021, replacing an earlier 2017 policy [9]. It sorts conditions into three groups: disorders curable with a one-time treatment such as transplantation; diseases needing long-term treatment at comparatively lower cost; and diseases for which definitive but very expensive lifelong therapy exists, where patient selection is difficult [9]. Financial assistance of up to Rs 50 lakh per patient is available for treatment at a notified Centre of Excellence (CoE) [9,10].

The CoE network has grown in stages. A Lok Sabha reply in July 2022 referred to eight CoEs [9]; by late 2024 there were twelve, with 63 rare diseases covered and 876 patients treated under the policy since its launch [10]. The same reply reported 7,414 patients with identified rare diseases recorded on the ICMR-maintained national registry portal [10]. By March 2025 the government reported thirteen CoEs, around 1,000 patients registered at them, and budget allocations of Rs 118.82 crore for 2024-25 and Rs 299.59 crore for 2025-26 [11]. At a national conference in May 2026, officials stated that the number of CoEs had reached fifteen, including two in the North-East, and that about 1,800 patients had received support under the policy; the Director General of the Indian Council of Medical Research acknowledged that the Rs 50 lakh ceiling may be insufficient given therapy costs [12].

These numbers reveal a gap. Fewer than two thousand patients supported, against a likely burden in the tens of millions if global prevalence estimates apply, points to under-diagnosis as much as under-funding. Utilisation has also been uneven: an analysis of ministry data in 2024 reported that the CoEs had used about 48% of roughly Rs 109 crore disbursed over three years, with wide variation between centres [13]. Treatment funding, in other words, cannot

reach patients who have not been diagnosed. That is the specific opening for diagnostic AI in India, and it is also why tools must work for Indian patients rather than merely being available to them. The point recurs in Sections 4.4 and 9.

3. Making Phenotypes Computable

3.1 The Human Phenotype Ontology

Before an algorithm can compare a patient with a disease, both must be described in the same language. Free-text clinical notes do not provide this. One clinician writes “wide-set eyes”, another “hypertelorism”, a third “increased interpupillary distance”. The **Human Phenotype Ontology (HPO)**, first described in 2008, solves the problem by providing a controlled, hierarchically organised vocabulary of phenotypic abnormalities, each with a stable identifier, and by annotating diseases with the terms that characterise them [14]. Because terms are arranged as a directed acyclic graph, a patient annotated with a specific term (for example, a particular type of cleft) is implicitly also annotated with its more general ancestors.

The ontology has grown steadily. The 2024 update reported 2,239 new terms added since late 2020, more than 49,000 new disease annotations, annotations covering over 8,100 rare diseases, and translations of term labels into ten languages besides English [15]. The translations are not cosmetic. A clinician in Maharashtra or Tamil Nadu who records findings in a local language, or who works with families who describe symptoms in one, needs a route into the same structured vocabulary that the diagnostic algorithms use, and extending translation to Indian languages is work that Indian clinical genetics groups are well placed to lead.

The HPO is one of a family of linked resources. Orphanet provides the nomenclature, classification and epidemiological data for rare diseases that underpin prevalence work such as that described in Section 2 [5]. Online Mendelian Inheritance in Man (OMIM) catalogues genes and genetic phenotypes and supplies many of the gene-disease links on which variant prioritisation depends. In practice the diagnostic tools discussed below draw on several of these at once.

3.2 Semantic similarity as the first machine-learning layer

Once phenotypes are coded, the question “which known disease does this patient most resemble?” becomes a computation. The Phenomizer, published in 2009, ranked candidate diseases by semantic similarity between the patient’s HPO terms and each disease’s annotation profile, weighting rare and specific features more heavily than common ones [16]. The idea is simple and has lasted. A patient with intellectual disability and seizures matches hundreds of conditions; a patient who also has a specific skeletal sign matches very few, and the specific sign should therefore count for more.

Later methods placed this reasoning on a firmer statistical footing. LIRICAL, for instance, reframed phenotype matching as a likelihood-ratio problem, so that each observed (or explicitly excluded) feature shifts the post-test probability of each candidate diagnosis in an interpretable way [17]. Interpretability matters here more than in many AI applications. A geneticist who must explain a diagnosis to a family, and defend it to colleagues, needs to know why a gene was ranked first.

Computable phenotypes also enable data exchange. The Global Alliance for Genomics and Health (GA4GH) Phenopacket schema defines a standard, machine-readable format for an individual's phenotypic and genetic findings [18]. Phenopackets are the connective tissue between hospital systems, research cohorts and matchmaking services. They are dull to describe and indispensable in practice.

The weak point of the whole edifice is input quality. HPO-based tools perform only as well as the terms entered. Busy clinicians record few terms, often generic ones, and rarely record features that were looked for and found absent. Automated extraction of HPO terms from clinical notes using natural language processing (NLP) is one response; Section 6.2 describes a pipeline in which NLP extraction matched an average of 4.3 relevant features per child, against 0.9 by manual review [1].

4. Facial Phenotyping with Deep Learning

4.1 From gestalt to algorithm

Many genetic syndromes produce a characteristic facial appearance. Experienced dysmorphologists recognise these patterns, often before they can articulate the individual features involved, which is why the field speaks of a facial "gestalt". The skill is scarce and slow to acquire. It is also exactly the kind of visual pattern recognition at which convolutional neural networks excel.

The landmark study is DeepGestalt, published by Gurovich and colleagues in 2019 and commercialised in the Face2Gene application [19]. The network was trained on 17,106 images of 10,953 individuals representing 216 syndromes. On a clinical test set of 502 images covering 92 syndromes, the correct syndrome appeared among the top ten suggestions in 90.6% of cases, and on a separate set of 329 published images in 89.4% [19]. In a more demanding experiment, the network distinguished five genetic subtypes of Noonan syndrome with 64% top-one accuracy, against 20% expected by chance [19]. That result suggested that facial appearance carries some information about the gene involved, not only about the syndrome.

Top-ten accuracy needs careful interpretation. It means the right answer is somewhere on a ranked list of ten, not that the system made the diagnosis. In practice its value is as a prompt:

a clinician who had not considered a particular syndrome is reminded of it, and confirmatory genetic testing follows.

4.2 What happens when the patient is unaffected

Most early evaluations tested DeepGestalt only on people known to have a syndrome. A 2020 diagnostic accuracy study by Pantel and colleagues asked the obvious next question: what does the system say about people who do not? They compared 323 individuals with 17 syndromes against controls matched for age, sex and ethnicity [20]. Top-ten sensitivity in affected individuals was 91%, consistent with earlier work. But the raw output separated syndromic from control faces only modestly, with an area under the receiver operating characteristic curve (AUROC) of 0.72; a support vector machine trained on the output raised this to 0.89 [20]. More strikingly, syndrome suggestions for unaffected individuals were far from random. Fragile X syndrome was proposed for 82% of control images, and the authors recommended that users be shown the false-positive rate of each suggested syndrome [20].

This is a familiar pattern in clinical AI. A system validated in the population it was built for looks excellent; tested in the population where it will actually be used, which includes many people without the target condition, its limitations appear. Face2Gene is best understood as a ranking aid within a genetics consultation, not as a screening tool.

4.3 GestaltMatcher and the ultra-rare tail

DeepGestalt is a classifier: it can only suggest syndromes it was trained on, and it needs a reasonable number of training images per syndrome. That rules out much of the long tail described in Section 2.1. GestaltMatcher, published in 2022, takes a different approach [21]. Instead of assigning a face to a fixed list of classes, it maps each face into a “clinical face phenotype space” in which distance reflects syndromic similarity. A new patient can then be compared with any previously seen patient, including those with disorders that have only a handful of known cases, or none yet in the literature.

Trained and tested on photographs of 17,560 patients with 1,115 rare disorders, GestaltMatcher separated 299 “frequent” syndromes, used to train the encoder, from 816 “rare” ones used for testing [21]. For the rare syndromes, top-ten accuracy was 17.67%, compared with 11.73% for an encoder trained only on healthy faces; for frequent syndromes it was 46.85% [21]. These figures are much lower than DeepGestalt’s, but the task is far harder: matching to more than a thousand possible disorders, most of which the encoder never saw during training. The practical demonstration was more persuasive than the benchmark. Across 15 gene-discovery studies, the method connected 40 of 79 families within the top 30 ranks, including families from institutes in India, Qatar, the United States and Switzerland [21]. Facial analysis had earlier contributed to the delineation of a new disorder, a LEMD2-associated nuclear envelopathy with early progeroid appearance [22].

4.4 Ancestry, bias and the Indian patient

The most serious limitation of facial phenotyping is not accuracy in the abstract but accuracy for whom. Dysmorphic features are judged against a background of normal facial variation, and that background differs between populations. One study showed that recognition of facial dysmorphism depended on the ethnic background of both the patient and the evaluating clinician [23]. Algorithms inherit the same problem through their training data. The DeepGestalt paper described its training set as approximately 43% Caucasian and 16% of other ethnicities, with the remainder unreported, and it did not analyse performance by ethnic group [19]. GestaltMatcher's authors acknowledged that facial similarity is higher within the same ethnicity and described this as a known confounder [21].

Several population-specific evaluations followed. Face2Gene was assessed in Japanese patients with congenital dysmorphic syndromes [24], in Thai children with Down syndrome [25] and in patients with Cornelia de Lange syndrome [26]. Of particular relevance to this volume, Narayanan and colleagues evaluated computer-aided facial analysis in Indian children with dysmorphic syndromes [27]. Such studies are valuable precisely because they test the tools where they are used, and they also generate images that can improve future training sets. The structural problem remains. A tool trained largely on one population and marketed globally will tend to perform best for that population, and the patients most poorly served will often be those who also have least access to expert dysmorphologists. Table 7.1 summarises the main evidence.

Table 7.1: Selected evaluations of deep-learning facial phenotyping in rare genetic disease

Study	Tool and approach	Data	Main result	Key limitation
Gurovich et al. 2019 [19]	DeepGestalt (Face2Gene); CNN classifier	17,106 training images, 216 syndromes	Top-10 accuracy 90.6% (clinical set, 502 images); 64% top-1 for 5 Noonan genotypes	Training set ~43% Caucasian; no ethnicity-stratified analysis
Pantel et al. 2020 [20]	DeepGestalt tested with matched unaffected controls	323 affected, 323 controls, 17 syndromes	Top-10 sensitivity 91%; AUROC 0.72 for affected vs control (0.89 with SVM)	Fragile X suggested for 82% of controls
Hsieh et al. 2019 [3]	PEDIA; facial scores combined with HPO and CADD for exome ranking	679 individuals, 105 monogenic disorders	Disease gene in top 10 in 99% with photos, vs <45% with CADD alone	Retrospective; cohort enriched for facial syndromes
Hsieh et al. 2022 [21]	GestaltMatcher; clinical face phenotype space for matching	17,560 patients, 1,115 disorders	Top-10 17.67% for unseen rare disorders; 40 of 79 families linked in gene-discovery studies	Ethnicity acknowledged as confounder
Population-specific evaluations [24,25,27]	Face2Gene in Japanese, Thai and Indian cohorts	Single-centre cohorts	Utility reported outside European training populations	Small samples; heterogeneous designs

5. Genomic Diagnosis and Variant Prioritisation

5.1 What sequencing achieves

The arrival of clinical exome sequencing in the early 2010s changed diagnostic genetics. Rather than test candidate genes one at a time, the laboratory sequences all of them and asks which variant explains the phenotype. A meta-analysis of studies in children with suspected genetic disease, covering 37 studies and 20,068 children, estimated pooled diagnostic utility of 0.41 (95% CI 0.34 to 0.48) for genome sequencing, 0.36 (0.33 to 0.40) for exome sequencing and 0.10 (0.08 to 0.12) for chromosomal microarray [28]. The difference between genome and exome sequencing was not statistically significant in that analysis, but both were far ahead of microarray.

The largest health-system evaluation is the pilot of the 100,000 Genomes Project in the English National Health Service. Among 4,660 participants from 2,183 families with 161 disorders, genome sequencing produced a diagnosis in 25% of probands [8]. Yield was 35% for disorders judged likely to be monogenic and 11% for those likely to be complex, and reached 40 to 55% for intellectual disability, hearing disorders and vision disorders [8]. Of the diagnoses, 14% came from a combination of research and automated approaches beyond the standard clinical pipeline [8]. Table 7.2 summarises these and other yield studies.

Table 7.2: Diagnostic yield of genomic testing in selected studies

Study	Setting and population	Test	Diagnostic yield	Notes
Clark et al. 2018 [28]	Meta-analysis, 37 studies, 20,068 children	WGS / WES / CMA	41% / 36% / 10% (pooled)	WGS vs WES difference not significant
Smedley et al. 2021 [8]	100,000 Genomes pilot, NHS England; 4,660 participants	Genome sequencing with Exomiser-assisted analysis	25% overall; 35% likely monogenic; 40-55% ID, hearing, vision	25% of diagnoses had immediate clinical consequences
Meng et al. 2017 [29]	278 critically ill infants, Texas	Clinical exome	36.7% overall; 50.8% for critical trio exome	Management changed in 52.0% of diagnosed infants
Farnaes et al. 2018 [30]	42 acutely ill infants, regional children's hospital	Rapid WGS	43% vs 10% by standard tests	Management changed in 31%
Clark et al. 2019 [1]	Seriously ill children, automated rWGS pipeline	Rapid WGS + NLP + automated interpretation	3 of 7 prospective ICU infants diagnosed	Retrospective concordance 97% recall, 99% precision

A yield of 25 to 40% is a remarkable advance on what came before, and it is also a reminder that most patients tested are not diagnosed. Some carry variants in genes not yet linked to disease; some have causal variants in non-coding regions or structural rearrangements that standard pipelines miss; some have causal variants that were found but could not be classified with confidence. Each of these failure modes is a target for machine learning.

5.2 The interpretation framework

The 2015 joint recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) remains the basis of clinical variant interpretation [31]. It replaced older terms such as "mutation" and "polymorphism" with five categories: pathogenic, likely pathogenic, uncertain significance, likely benign and benign. "Likely" was defined as greater than 90% certainty [31]. Evidence is weighed through a set of criteria covering population frequency, segregation, functional data, de novo occurrence and others.

Computational prediction enters this framework through a narrow door. Criterion PP3 allows "multiple lines of computational evidence" supporting a deleterious effect to count as supporting evidence of pathogenicity, and BP4 allows the converse for benign classification; each may be applied only once per variant, because many prediction tools draw on similar inputs and would otherwise be double-counted [31]. In 2015 this caution was sensible. As Section 5.4 explains, newer predictors are considerably more accurate than those available when the guidelines were written, and how much weight they should carry is an active question.

The category that causes most trouble is the middle one. A **variant of uncertain significance (VUS)** is one for which evidence is insufficient or conflicting. It should not be used for clinical decisions, yet it is reported, and it often causes anxiety and confusion for families. Public databases such as ClinVar aggregate classifications submitted by laboratories worldwide and expose conflicts between them, assigning each record a review status [32]. The VUS problem is partly a data problem: many variants are uncertain simply because they have been seen too rarely, or only in populations that are poorly represented in reference databases. That second cause bears directly on Indian patients, whose variants are more likely to be absent from predominantly European reference sets.

5.3 Phenotype-driven prioritisation: Exomiser

The core computational problem of genomic diagnosis is ranking. A typical exome contains more than 30,000 variants, and standard filtering for quality, population frequency and predicted consequence usually leaves up to about 1,000 candidates [2]. A genome contains several million variants; one clinical pipeline described the task as finding the causal change among about 4.5 million [33]. Figure 7.2 shows the narrowing schematically. No analyst can review a thousand variants for each of thousands of patients. The answer has been to combine variant-level evidence with gene-level phenotype evidence.

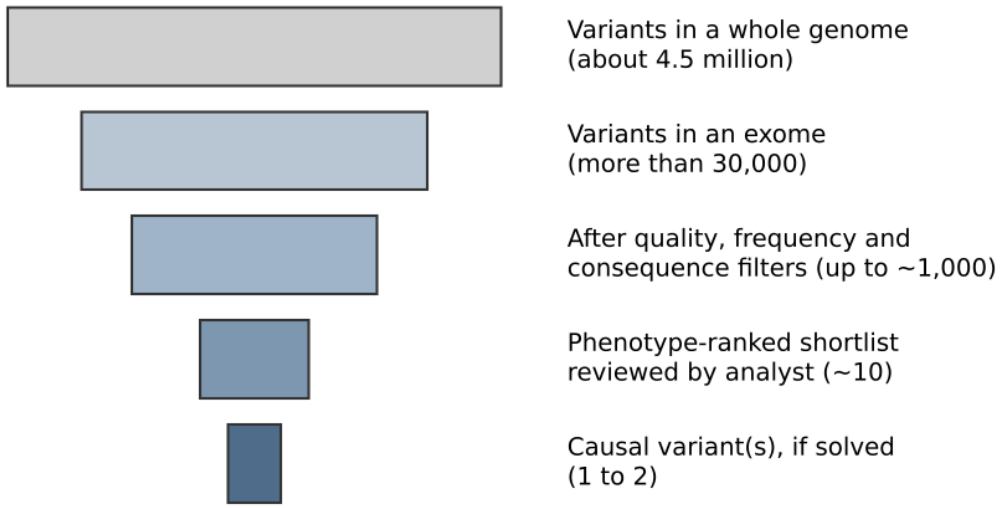

Figure 7.2: Variant prioritisation funnel from whole genome to causal variant. Counts for genome, exome and post-filter stages are taken from published descriptions [2,33]; the shortlist and final stages are illustrative.

Exomiser, developed by an international academic collaboration, is the most widely used open tool of this kind [2]. It annotates and filters variants, then scores each candidate gene by how well the patient's HPO profile matches phenotypes associated with that gene in human disease databases, in mouse and zebrafish models, and through neighbouring genes in protein interaction networks [2,34]. The cross-species step matters because many human genes have no known disease association but do have model-organism phenotypes. In simulation, the combined hiPHIVE method ranked the correct variant first in 97% of samples involving known disease genes [2]. Adding mouse, zebrafish and interaction data raised detection of novel disease-gene associations from 45% with human data alone to 74% [35]. The software runs on a desktop computer, analysing a single exome in seconds to a minute or two [2].

A worked example in the Exomiser protocol makes the point well. For a patient with Pfeiffer syndrome, the causal gene *FGFR2* ranked first when phenotype data were supplied; without them, it ranked 31st [2]. Phenotype information, not variant information alone, does most of the work in the final step.

The 100,000 Genomes Project made Exomiser central to a national health service. Genomics England uses it as the automated variant prioritisation framework for all rare disease referrals, taking as input the family's variant calls, the pedigree and the proband's HPO terms [36]. The pilot found that research and automated approaches together contributed 14% of diagnoses [8].

The integration of tools is also visible in PEDIA, which combined DeepGestalt facial scores with HPO-based scores and CADD variant scores. In 679 individuals with 105 monogenic disorders, the disease gene was ranked in the top ten in under 45% of cases by CADD alone, in 63 to 94% when phenotype similarity was added, and in 99% when a frontal photograph was also included [3]. Earlier, Zemojtel and colleagues had shown in a targeted-panel setting that computational phenotype analysis could guide diagnosis effectively [37].

These results are impressive but mostly retrospective, and the cohorts were assembled from patients already diagnosed. Real-world performance is lower, because real patients have incomplete phenotyping, atypical presentations and causal variants in genes the algorithms cannot score well.

5.4 Machine-learned variant-effect predictors

Variant-level scoring is the second input to prioritisation. Older tools such as SIFT and PolyPhen relied mainly on sequence conservation and protein features. The tools now dominant in clinical pipelines are machine-learned and broader in scope (Table 7.3).

CADD (Combined Annotation-Dependent Depletion), introduced in 2014, took an unusual approach to the shortage of labelled pathogenic variants. Rather than train on curated disease mutations, it trained a support vector machine to distinguish 14.7 million variants that have become fixed in humans since the split from the great apes (and are therefore mostly tolerated) from 14.7 million simulated variants (which include many deleterious ones) [38]. Scores were precomputed for all 8.6 billion possible single-nucleotide variants in the human genome [38]. Later versions integrated more than 60 annotations, added the GRCh38 genome build and recommended that scores be used for ranking rather than with a single universal cutoff [39]. That recommendation is worth heeding. Many laboratories nonetheless apply fixed thresholds.

SpliceAI, published in 2019, addressed a class of variants that conventional annotation largely ignored. A deep neural network reads the surrounding pre-mRNA sequence and predicts whether each position is a splice donor or acceptor, and how a variant alters that prediction [40]. Variants predicted with high confidence validated on RNA sequencing at about three-quarters the rate of disruptions of the canonical splice dinucleotides, and the authors estimated that cryptic splice mutations account for roughly 11% of pathogenic mutations in autism spectrum disorder and 9% in intellectual disability [40]. For rare disease diagnosis, the importance is that synonymous and intronic variants, routinely filtered out as irrelevant, can now be flagged for review.

AlphaMissense, a deep-learning model released by Google DeepMind in 2023, extended variant-effect prediction to missense variants across the whole proteome [41]. It produced predictions for all possible single amino acid substitutions in human proteins and classified

89% of missense variants as likely benign or likely pathogenic [41]. Since missense variants make up a large share of VUS, this attracted considerable attention.

The caution is in the vocabulary. A predictor's "likely pathogenic" label is not the ACMG category of the same name. Under the 2015 framework, computational evidence contributes at most a supporting criterion [31]. A VUS cannot be reclassified as pathogenic on the strength of a prediction alone, however accurate the model is on benchmarks, because benchmark sets themselves are drawn from variants that have already been classified, often partly with the help of earlier predictors. There is a circularity risk here, and it is sharpest for variants in under-studied genes and under-represented populations, the very variants for which help is most needed. Predictors are best used as they are designed to be used in Exomiser and similar tools: to order candidates for human review.

Table 7.3: Machine-learning tools used in rare disease diagnosis

Tool	Input	Method	Validation reported	Principal limitation
Phenomizer [16]	HPO terms	Ontology semantic similarity	Methods paper introducing the approach	Depends on quality and specificity of terms
LIRICAL [17]	HPO terms and variants	Likelihood-ratio ranking	Methods paper; interpretable per-feature output	Same input dependence; less suited to novel genes
DeepGestalt / Face2Gene [19]	2D facial photograph	Deep CNN classifier	Top-10 90.6% on clinical test set	Fixed syndrome list; ancestry imbalance
GestaltMatcher [21]	2D facial photograph	Learned embedding space, nearest-neighbour matching	Top-10 17.67% for unseen rare disorders	Low absolute accuracy; ethnicity confound
Exomiser [2]	VCF, pedigree, HPO terms	Variant filtering plus cross-species phenotype and network scoring	Top-ranked in 97% of simulations (known genes)	Simulations optimistic; real-world yield lower
CADD [38,39]	Any SNV or short indel	SVM, later logistic regression, on >60 annotations	Ranks known pathogenic variants highly	Not a calibrated probability; misuse of cutoffs
SpliceAI [40]	Pre-mRNA sequence context	Deep neural network	RNA-seq validation of predicted cryptic splice sites	Predictions still need RNA-level confirmation
General-purpose LLMs [42]	Case description or HPO terms	Large language model	Top-1 23.6% (best model) vs 35.5% for Exomiser on 5,213 cases	Below established tools; training-data contamination risk
AlphaMissense [41]	Missense variant	Deep-learning model, proteome-wide	Classifies 89% of missense variants	Only supporting evidence under ACMG/AMP; circularity risk

6. Rapid Genome Sequencing in Critical Care

6.1 Why speed matters in the NICU

Genetic disease accounts for a substantial share of serious illness in neonatal and paediatric intensive care [43]. In this setting a diagnosis that arrives in weeks may arrive too late. The infant may have died, been subjected to unnecessary procedures, or missed a window in which a specific therapy could have prevented permanent harm. Rapid whole-genome sequencing was developed to meet this need, and it is one of the clearest cases in which automation, including AI, has directly changed what clinicians can do.

The first demonstration, published in 2012, achieved a differential diagnosis in about 50 hours from consent to preliminary report [44]. That pipeline already relied on software, called SSAGA, which mapped clinical features to candidate genes to narrow the search [44]. Subsequent studies established clinical value. Among 278 critically ill infants tested by clinical exome at a Texas centre, 36.7% received a diagnosis, rising to 50.8% with a rapid "critical trio" exome that had a median turnaround of 13 days [29]. In a cohort of 42 acutely ill infants at a regional children's hospital, rapid genome sequencing diagnosed 43% compared with 10% by standard tests, and changed management in 31% [30].

6.2 Automation and the race against time

The Rady Children's Institute for Genomic Medicine in San Diego pushed the timeline further by automating two steps that had previously required human experts. NLP software extracted phenotype terms directly from the electronic health record, with 80% precision and 93% recall against manual review [1]. Automated interpretation software then ranked candidate variants. Retrospectively, in 95 children with 97 genetic diagnoses, the automated system agreed with expert interpretation with 97% recall and 99% precision [1]. The median time from sample to provisional diagnosis was 20 hours 10 minutes [1]. Commercial components included an NLP platform for record abstraction and an AI interpretation engine that could propose a causal variant among roughly 4.5 million in about five minutes [33].

One case from that programme shows what the speed can mean. A five-week-old boy with encephalopathy and seizures underwent rapid genome sequencing, and a provisional diagnosis of thiamine metabolism dysfunction syndrome 2, caused by a frameshift variant in the thiamine transporter gene *SLC19A3*, was made 14 hours 33 minutes after the blood sample reached the laboratory [45]. Treatment with thiamine and biotin followed. After treatment, only one brief seizure was recorded, and the child was reported to be thriving at seven months [45]. His parents were first cousins, and a previous sibling with a similar illness had died at eleven months without a diagnosis [45]. The contrast between the two children is the clearest argument for rapid diagnosis that the literature contains.

In 2022 Gorzynski and colleagues reported ultrarapid genome sequencing using nanopore technology, with a fastest time of 7 hours 18 minutes from sample to diagnosis; five of twelve patients were diagnosed [46]. Figure 7.3 sets these milestones side by side.

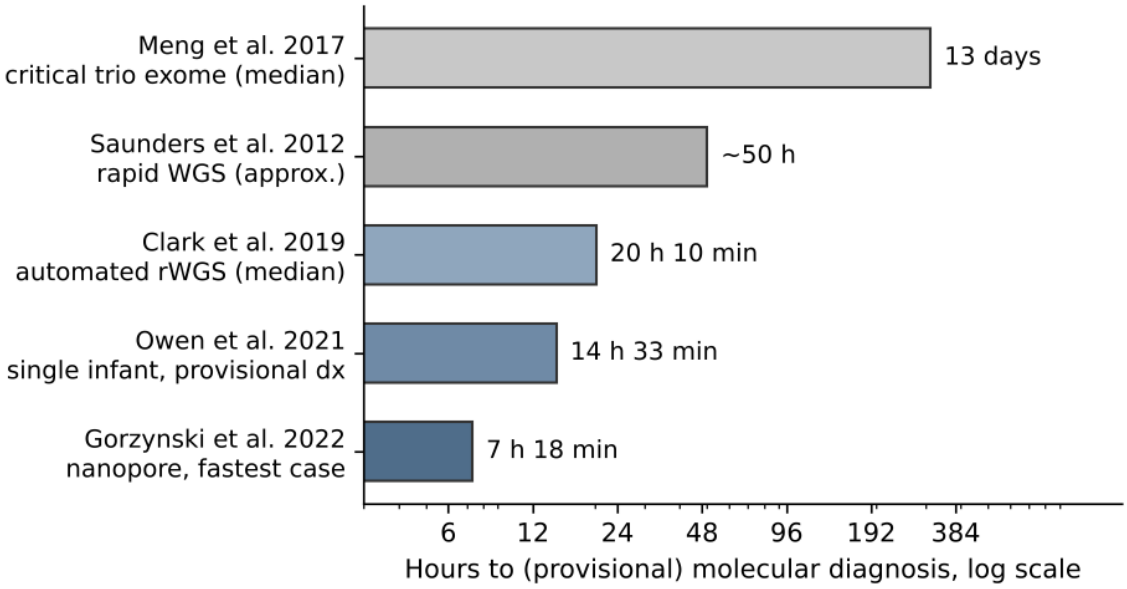

Figure 7.3: Reported times to molecular diagnosis in critical-care genomic sequencing studies, log scale. Endpoints differ between studies (median, approximate or single fastest case) and are not directly comparable [1,29,44-46].

The figure invites a caution. These are records and medians from well-funded academic centres, not typical service turnaround times, and fastest-case figures say little about throughput. Rapid sequencing also depends on a chain of conditions that most hospitals in India and other low- and middle-income countries do not yet meet: on-site or nearby sequencing capacity, round-the-clock bioinformatics, clinicians trained to order and interpret the test, and funding for a test whose cost is front-loaded while its savings are spread over the subsequent course of care. AI shortens the analytical step. It does not remove the others.

7. Beyond the Sequence: Health Records, Language Models and Screening

7.1 Finding the undiagnosed in health records

Most patients with an undiagnosed rare disease are not in a genetics clinic. They are in general paediatrics, neurology, nephrology or internal medicine, often labelled with a symptomatic diagnosis. A natural idea is to search electronic health records for the combinations of features, encounters and investigations that suggest an underlying genetic condition, and flag those

patients for genetic evaluation. The NLP work described in Section 6.2 shows that the extraction step is feasible within a single institution [1]. One of the more rigorous examples used diagnostic billing codes rather than free text. Trained on 2,286 patients who had undergone chromosomal microarray testing and 9,144 matched controls, the model reached an AUROC of 0.97 in held-out data and 0.95 in an independent hospital system, yet the area under the precision-recall curve fell from 0.92 to 0.62 in the external system [47]. The drop in precision, with discrimination almost unchanged, shows how strongly setting and prevalence shape the yield of any flagging tool.

Even so, the evidence for population-level case-finding is thinner than the enthusiasm around it. Many published models target one disease at a time, are developed and tested in a single health system, and report retrospective discrimination rather than prospective diagnostic yield. The base-rate problem is severe. For a condition affecting one in fifty thousand people, even a model with excellent specificity will generate many false alarms for each true case, and every flag carries a cost in clinician time, testing and patient worry. Health records also encode the behaviour of the health system: patients who have been referred and investigated more have richer records, so models can learn to find patients who are already on the way to diagnosis. In India, where most outpatient care is not captured in structured electronic records, and where the Ayushman Bharat Digital Mission (discussed in Chapter 10) is still building the infrastructure, record-based case-finding is a longer-term prospect.

7.2 Large language models

General-purpose large language models (LLMs) have attracted interest as differential diagnosis aids, since they have absorbed much of the published medical literature, including case reports of rare conditions. The most systematic comparison so far encoded 5,213 published case reports as phenopackets and set several LLMs against Exomiser. The best-performing model, o1-preview, ranked the correct diagnosis first in 23.6% of cases and within the top ten in 36.8%; Exomiser achieved 35.5% and 58.5% [42]. The authors concluded that LLMs had not yet reached the accuracy of established decision-support tools. That finding should be read alongside three further cautions.

The first is contamination. Published case reports, the most convenient test material, may have been part of the model's training data, so apparent diagnostic skill may partly reflect recall. The second is the absence of a genetic anchor. An LLM ranking diagnoses from a vignette is doing something closer to the Phenomizer than to Exomiser; it cannot confirm a diagnosis, and its value lies in widening the differential before testing. The third is fluency without calibration. LLMs produce confident, well-written text whether or not the content is correct, and they can generate plausible but non-existent gene-disease associations or references. For a clinician

without specialist genetics training, which describes most clinicians who first see a rare disease patient, that combination is hazardous.

The more defensible near-term roles are narrower: converting free-text notes into HPO terms for input to established tools, summarising long records before a genetics consultation, and drafting family-facing explanations that a clinician then checks. Each of these keeps the LLM upstream of a validated system or downstream of a human expert. In preparing this chapter (November 2026), no prospective evidence was identified that LLM-based differential diagnosis improves time to diagnosis or diagnostic yield for rare disease patients in routine care, and claims to that effect should be treated as unproven.

7.3 Newborn screening

Newborn screening by biochemical methods, mainly tandem mass spectrometry, is the oldest population-scale rare disease programme, and it already involves algorithmic decision-making in the cut-offs that separate screen-positive from screen-negative results. Machine learning can use the full metabolite profile, rather than a few cut-offs, to separate true from false positives. In data from 2,777 screen-positive infants in the California programme, 235 of whom had a confirmed disorder, a machine-learning classifier reduced false positives by 89% for glutaric acidemia type 1, 98% for ornithine transcarbamylase deficiency and 45% for methylmalonic acidemia without loss of sensitivity, but by only 2% for very long-chain acyl-CoA dehydrogenase deficiency [48]. The gain is real and uneven, and depends on the disorder. Genome-based newborn screening, in which a defined panel of treatable conditions is analysed from sequence, is being piloted in several high-income countries and would depend heavily on automated interpretation of the kind described in Section 5.

For India, the more pressing question is the reach of screening itself rather than the sophistication of its algorithms. National policy statements have emphasised genetic analysis, counselling and screening, and officials in 2026 cited the UMMID programme for inherited disorders and its NIDAN Kendras as channels through which patients have been supported [12]. Machine learning can refine a screening programme; it cannot substitute for one.

8. AI for Treatment: Knowledge Graphs and Drug Repurposing

Diagnosis is the first problem; treatment is the second. For most rare diseases there is no approved therapy, and the economics of conventional drug development rarely favour conditions with a few hundred patients worldwide. Drug repurposing, the identification of new uses for existing drugs, is attractive because safety data already exist and development costs are lower [49].

Knowledge graphs provide a natural substrate for computational repurposing. They encode drugs, diseases, genes, proteins, phenotypes and pathways as nodes, with edges representing

known relationships such as a drug targeting a protein or a gene being associated with a disease [50]. A model can then learn patterns in the graph and predict missing edges, such as an unrecorded therapeutic relationship between a drug and a disease.

TxGNN, published in 2024, is the most prominent rare disease example [51]. It is a graph neural network trained on a large medical knowledge graph and designed for zero-shot prediction, that is, ranking drugs for diseases that have no known treatments in the training data. It covers 17,080 diseases, about 92% of which have no approved drug indication, and 7,957 candidate drugs [51]. Under zero-shot evaluation the authors reported gains of 49.2% for indication prediction and 35.1% for contraindication prediction relative to benchmark methods [51]. The model borrows strength from diseases that share mechanisms or phenotypes, which is the only way such predictions can be made for diseases with almost no data of their own. The authors tested whether top-ranked predictions corresponded to real-world off-label prescribing in records from more than 1.2 million patients at a single US health system, and evaluated an explanation interface with a small group of clinicians and pharmacists [51].

These are hypothesis-generating results. Agreement with off-label prescribing shows that the model recovers what clinicians already do; it does not show that the novel predictions work. For a rare disease with no treatment, the only real test is a clinical study, and such studies are themselves hard to run in small populations. In India, where the NPRD Group 3 conditions include lysosomal storage disorders with very expensive enzyme therapies, and where the ICMR has supported work on repurposed and indigenous therapies [12], computational repurposing could help prioritise which candidates merit a trial. It is a filter for hypotheses, not a source of evidence.

9. Cross-Cutting Challenges

9.1 Small data and the few-shot problem

The defining technical difficulty of rare disease AI is that the categories of interest have very few examples. Deep learning works best with thousands of labelled examples per class; many rare diseases have fewer than a hundred known patients worldwide, and some have one family. Three broad strategies have emerged.

The first is to learn a general representation from abundant data and apply it to scarce data. GestaltMatcher does this by learning a face space from more common syndromes and then matching ultra-rare ones by distance [21]; TxGNN does it by learning from treated diseases and transferring to untreated ones [51]. The second is to encode prior knowledge so that the algorithm need not learn it. Ontologies such as the HPO, and cross-species phenotype data in Exomiser, inject decades of curated biology into the model [15,34]. The third is to combine

weak signals: PEDIA showed that face, phenotype terms and variant scores together far outperform any one of them [3].

Each strategy has a characteristic weakness. Transfer works only when the new disease resembles the old ones in the learned space. Curated knowledge is incomplete and biased toward well-studied genes, which means that algorithms built on it tend to rediscover what is already known. Combining signals requires all inputs to be available, and photographs and detailed phenotypes are often missing.

9.2 Data sharing and matchmaking

For a newly recognised genetic disorder, the key evidence is often a second unrelated family with a similar phenotype and a damaging variant in the same gene. Finding that family is a search problem across institutions and countries. The Matchmaker Exchange, launched in 2015, created a federated network through which databases such as GeneMatcher and PhenomeCentral can be queried with a gene, a phenotype profile or both, without pooling raw data centrally [4,52,53]. The GestaltMatcher gene-discovery results described in Section 4.3 relied on GeneMatcher collaborations [21].

Matchmaking is AI-adjacent rather than AI in the narrow sense, but it shows what rare disease machine learning needs most: data that can be found and compared across borders. Standards such as Phenopackets make records interoperable [18]; federated approaches (Chapter 8) allow models to learn across sites without transferring identifiable genomes. India's Digital Personal Data Protection Act 2023 and the ethics of sharing genomic and facial data, which are among the most identifiable forms of personal information, will shape how Indian centres participate. Facial photographs in particular cannot be meaningfully anonymised.

9.3 Equity

Equity concerns run through every section of this chapter. Facial algorithms trained mostly on European faces [19,23]; reference databases in which South Asian variation is thinly represented, so that more variants found in Indian patients remain uncertain; rapid sequencing concentrated in a few wealthy hospitals [46]; and treatment funding that cannot reach the undiagnosed [12,13]. None of these is solved by better algorithms alone. Some are partly addressable by AI itself, for example by explicit evaluation of performance in under-represented groups and by training on more diverse data. Most require investment in sequencing capacity, genetic counselling, clinician education and data infrastructure. Table 7.4 summarises the main challenges and plausible responses.

Table 7.4: Principal challenges for AI in rare disease diagnosis and possible mitigations

Challenge	Why it matters	Possible mitigation	Evidence or example
Very few cases per disease	Classifiers cannot learn ultra-rare classes	Embedding and matching, transfer, zero-shot designs	GestaltMatcher [21]; TxGNN [51]
Poor phenotype capture	Ranking tools depend on HPO input	NLP extraction from records; Phenopackets	4.3 vs 0.9 matched features per child [1]; [18]
False positives in unaffected people	Tools validated only in affected cohorts	Test against matched controls; report per-syndrome false-positive rates	Fragile X suggested in 82% of controls [20]
Ancestry imbalance	Lower accuracy and more VUS for under-represented groups	Population-specific evaluation and data	Japanese, Thai, Indian cohorts [24,25,27]; [23]
Over-reliance on predictors	Computational scores mistaken for classification	Keep predictors as PP3/BP4 supporting evidence; ranking not cutoffs	[31,39]
Fragmented data	Second families and new genes not found	Federated matchmaking; standards	Matchmaker Exchange [4]
Infrastructure and cost	Rapid sequencing confined to few centres	Regional hubs; CoE networks; funding linked to diagnosis	NPRD CoEs [11,12]

10. Conclusion and Future Directions

The most consistent lesson from a decade of work is that AI in rare disease diagnosis succeeds when it ranks and fails when it is asked to decide. Exomiser, PEDIA, DeepGestalt and the automated rapid-genome pipelines all reduce a search space so that a trained person can make a judgement. Where the evidence is strongest, as in the 100,000 Genomes Project and the NICU programmes, the tools are embedded in a workflow that ends with expert review and ACMG/AMP classification, as sketched in Figure 7.1. Where claims run ahead of evidence, as with some uses of language models and record-based case-finding, the missing piece is usually prospective validation in the population where the tool will be used.

Three priorities follow for the coming years. The first is diversity of data. Facial and genomic tools need training and evaluation data from South Asian, African and other under-represented populations, and Indian clinical genetics centres, including the NPRD CoEs, are well placed to contribute images, phenotypes and variant classifications under appropriate consent and governance. The second is honest evaluation: studies that include unaffected controls, report performance by ancestry and measure clinical endpoints such as time to diagnosis rather than retrospective ranking accuracy alone. The third is reanalysis. A negative genome today is not a negative genome forever. As gene-disease knowledge grows and predictors improve, systematic, automated reanalysis of unsolved cases could deliver new diagnoses at low marginal cost, provided the data are stored and the phenotypes kept current.

For India specifically, the arithmetic is stark. The NPRD has created a treatment pathway for a few thousand patients, while a burden in the tens of millions remains largely undiagnosed.

Diagnostic AI cannot close that gap by itself, but it can make scarce genetic expertise go further, provided the tools are validated in Indian patients and deployed inside a system that can act on what they find.

REFERENCES

1. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, 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
2. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10(12):2004-15. https://doi.org/10.1038/nprot.2015.124
3. Hsieh TC, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, et al. PEDIA: prioritization of exome data by image analysis. Genet Med. 2019;21(12):2807-14. https://doi.org/10.1038/s41436-019-0566-2
4. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat. 2015;36(10):915-21. https://doi.org/10.1002/humu.22858
5. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165-73. https://doi.org/10.1038/s41431-019-0508-0
6. EURORDIS-Rare Diseases Europe. Rare Barometer survey: time to diagnosis and determinants of diagnostic delays [Internet]. Paris: EURORDIS; 2024 May [cited 2026 Nov 3]. Available from: https://www.eurordis.org/rb-diagnosis-article
7. International Rare Diseases Research Consortium (IRDiRC). Vision and goals [Internet]. IRDiRC; [cited 2026 Nov 3]. Available from: https://irdirc.org/vision-and-goals/
8. 100,000 Genomes Project Pilot Investigators; Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000 Genomes pilot on rare-disease diagnosis in health care: preliminary report. N Engl J Med. 2021;385(20):1868-80. https://doi.org/10.1056/NEJMoa2035790
9. Press Information Bureau, Ministry of Health and Family Welfare, Government of India. National Policy for Rare Diseases, 2021 [written reply in Lok Sabha; press release]. New Delhi: PIB; 2022 Jul 29 [cited 2026 Nov 3]. Available from: https://pib.gov.in/PressReleasePage.aspx?PRID=1846230
10. Lok Sabha, Parliament of India. Unstarred Question No. 701: National Policy for Rare Diseases, 2021. Answered by the Minister of State for Health and Family Welfare; 2024 Nov 29 [cited 2026 Nov 3]. Available from: https://eparlib.sansad.in/bitstream/123456789/2983511/1/AU701_nbdruV.pdf

11. Lok Sabha, Parliament of India. Unstarred Question No. 4610: rare diseases. Answered by the Ministry of Health and Family Welfare; 2025 Mar 28 [cited 2026 Nov 3]. Available from: https://eparlib.sansad.in/bitstream/123456789/3014825/1/AU4610_2nhcja.pdf
12. Indian Pharma Post. Healthcare policymakers lay thrust on genetic screening and indigenous therapies for rare diseases. Indian Pharma Post [Internet]. 2026 May 5 [cited 2026 Nov 3]. Available from: https://www.indianpharmapost.com/public-health/lite/healthcare-policymakers-lay-thrust-on-genetic-screening-and-indigenous-therapies-for-rare-diseases-20071
13. Business Standard. Only 4 of 12 CoEs for rare diseases utilise NPRD funds in past 3 years. Business Standard [Internet]. 2024 Jul 28 [cited 2026 Nov 3]. Available from: https://www.business-standard.com/health/only-4-of-12-coes-for-rare-diseases-utilise-nprd-funds-in-past-3-years-124072800341_1.html
14. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83(5):610-5. https://doi.org/10.1016/j.ajhg.2008.09.017
15. Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res. 2024;52(D1):D1333-46. https://doi.org/10.1093/nar/gkad1005
16. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457-64. https://doi.org/10.1016/j.ajhg.2009.09.003
17. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Am J Hum Genet. 2020;107(3):403-17. https://doi.org/10.1016/j.ajhg.2020.06.021
18. Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat Biotechnol. 2022;40(6):817-20. https://doi.org/10.1038/s41587-022-01357-4
19. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60-4. https://doi.org/10.1038/s41591-018-0279-0
20. Pantel JT, Hajjir N, Danyel M, Elsner J, Abad-Perez AT, Hansen P, et al. Efficiency of computer-aided facial phenotyping (DeepGestalt) in individuals with and without a genetic syndrome: diagnostic accuracy study. J Med Internet Res. 2020;22(10):e19263. https://doi.org/10.2196/19263
21. Hsieh TC, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nat Genet. 2022;54(3):349-57. https://doi.org/10.1038/s41588-021-01010-x
22. Marbach F, Rustad CF, Riess A, Đukić D, Hsieh TC, Jobani I, et al. The discovery of a LEMD2-associated nuclear envelopathy with early progeroid appearance suggests advanced applications for AI-driven facial phenotyping. Am J Hum Genet. 2019;104(4):749-57. https://doi.org/10.1016/j.ajhg.2019.02.021

23. Lumaka A, Cosemans N, Lulebo Mampasi A, Mubungu G, Mvuama N, Lubala T, et al. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. Clin Genet. 2017;92(2):166-71. https://doi.org/10.1111/cge.12948
24. Mishima H, Suzuki H, Doi M, Miyazaki M, Watanabe A, Matsumoto T, et al. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. J Hum Genet. 2019;64(8):789-94. https://doi.org/10.1038/s10038-019-0619-z
25. Vorravanpreecha N, Lertboonnum T, Rodjanadit R, Sriplienchan P, Rojnueangnit K. Studying Down syndrome recognition probabilities in Thai children with de-identified computer-aided facial analysis. Am J Med Genet A. 2018;176(9):1935-40. https://doi.org/10.1002/ajmg.a.40483
26. Latorre-Pellicer A, Ascaso Á, Trujillano L, Gil-Salvador M, Arnedo M, Lucia-Campos C, et al. Evaluating Face2Gene as a tool to identify Cornelia de Lange syndrome by facial phenotypes. Int J Mol Sci. 2020;21(3):1042. https://doi.org/10.3390/ijms21031042
27. Narayanan DL, Ranganath P, Aggarwal S, Dalal A, Phadke SR, Mandal K. Computer-aided facial analysis in diagnosing dysmorphic syndromes in Indian children. Indian Pediatr. 2019;56(12):1017-9.
28. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med. 2018;3:16. https://doi.org/10.1038/s41525-018-0053-8
29. Meng L, Pammi M, Saronwala A, Magoulas P, Ghazi AR, Vetrini F, et al. Use of exome sequencing for infants in intensive care units: ascertainment of severe single-gene disorders and effect on medical management. JAMA Pediatr. 2017;171(12):e173438. https://doi.org/10.1001/jamapediatrics.2017.3438
30. Farnaes L, Hildreth A, Sweeney NM, Clark MM, Chowdhury S, Nahas S, et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. NPJ Genom Med. 2018;3:10. https://doi.org/10.1038/s41525-018-0049-4
31. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24. https://doi.org/10.1038/gim.2015.30
32. National Center for Biotechnology Information. What is ClinVar? [Internet]. Bethesda (MD): National Library of Medicine; updated 2026 Apr 29 [cited 2026 Nov 3]. Available from: https://www.ncbi.nlm.nih.gov/clinvar/intro/
33. Inside Precision Medicine. Rady researchers use AI pipeline for rapid diagnosis of rare genetic diseases [Internet]. [cited 2026 Nov 3]. Available from: https://www.insideprecisionmedicine.com/news-and-features/rady-researchers-use-ai-pipeline-for-rapid-diagnosis-of-rare-genetic-diseases
34. Robinson PN, Köhler S, Oellrich A; Sanger Mouse Genetics Project; Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340-8. https://doi.org/10.1101/gr.160325.113

35. Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet Med. 2016;18(6):608-17. https://doi.org/10.1038/gim.2015.137

36. Genomics England. Rare disease analysis guide: Exomiser overview [Internet]. Version 1.134. London: Genomics England; 2024 Mar 20 [cited 2026 Nov 3]. Available from: https://pipeline-rd-help.genomicsengland.co.uk/latest/variant-prioritisation-approaches/exomiser/overview/

37. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6(252):252ra123. https://doi.org/10.1126/scitranslmed.3009262

38. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5. https://doi.org/10.1038/ng.2892

39. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-94. https://doi.org/10.1093/nar/gky1016

40. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-48.e24. https://doi.org/10.1016/j.cell.2018.12.015

41. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381:eadg7492. https://doi.org/10.1126/science.adg7492

42. Reese JT, Chimirri L, Bridges Y, Danis D, Caufield JH, Gargano MA, et al. Systematic benchmarking demonstrates large language models have not reached the diagnostic accuracy of traditional rare-disease decision support tools. Eur J Hum Genet. 2026;34(4):498-504. https://doi.org/10.1038/s41431-026-02054-5

43. Petrikin JE, Willig LK, Smith LD, Kingsmore SF. Rapid whole genome sequencing and precision neonatology. Semin Perinatol. 2015;39:623-31. https://doi.org/10.1053/j.semperi.2015.09.009

44. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Abu Alnadi N, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med. 2012;4(154):154ra135. https://doi.org/10.1126/scitranslmed.3004041

45. Owen MJ, Niemi AK, Dimmock DP, Speziale M, Nespeca M, Chau KK, et al. Rapid sequencing-based diagnosis of thiamine metabolism dysfunction syndrome. N Engl J Med. 2021;384(22):2159-61. https://doi.org/10.1056/NEJMc2100365

46. Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al. Ultrarapid nanopore genome sequencing in a critical care setting. N Engl J Med. 2022;386(7):700-2. https://doi.org/10.1056/NEJMc2112090

47. Morley TJ, Han L, Castro VM, Morra J, Perlis RH, Cox NJ, et al. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. Nat Med. 2021;27(6):1097-104. https://doi.org/10.1038/s41591-021-01356-z

48. Peng G, Tang Y, Cowan TM, Enns GM, Zhao H, Scharfe C. Reducing false-positive results in newborn screening using machine learning. Int J Neonatal Screen. 2020;6(1):16. https://doi.org/10.3390/ijns6010016

49. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41-58. https://doi.org/10.1038/nrd.2018.168

50. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10:67.

51. Huang K, Chandak P, Wang Q, Havaldar S, Vaid A, Leskovec J, et al. A foundation model for clinician-centered drug repurposing. Nat Med. 2024;30(12):3601-13. https://doi.org/10.1038/s41591-024-03233-x

52. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat. 2015;36(10):928-30.

53. Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, et al. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum Mutat. 2015;36(10):931-40. https://doi.org/10.1002/humu.22851