AIVA: An Agentic Platform for Phenotype-Aware Variant Analysis, Interpretation and Clinical Decision Support in Rare Disease

Tarun Mamidi

tarun@mamidi.ai

Mamidi Health

Arun K. Boddapati

software

Keywords: variant interpretation, gene prioritization, rare disease diagnostics, agentic AI, large language models, ACMG/AMP classification, clinical genomics

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10286684/v1

Additional Declarations: Competing interest reported. Tarun Mamidi is the founder of Mamidi Health.

Abstract

Background

Rare disease diagnosis remains slow, with patients facing a diagnostic odyssey averaging 5 to 7 years. Genomic sequencing has shifted the bottleneck to interpretation: the challenge is identifying the pathogenic variant(s) among ranked candidates. Existing classifiers and phenotype-driven prioritization tools filter and rank candidates, but the review still occurs outside the platform, where analysts manually curate the literature and databases to weigh the evidence for pathogenicity. In addition, current automated classifiers apply generic ACMG/AMP rules rather than gene-specific expert-panel specifications, and most confine users to their own fixed annotations. We present AIVA (AI-powered Variant Analysis), an agentic platform that addresses this interpretive gap through a conversational interface, performing phenotype-aware prioritization and ACMG/AMP classification, grounding every assessment in tool-retrieved evidence with attached citations.

Results

Benchmarking variant classification on 8,387 ClinGen expert-panel-curated variants across 108 genes and 35 Variant Curation Expert Panels, AIVA reached a macro F1 of 80.5%, ahead of the rule-based classifiers BIAS-2015 (75.3%) and InterVar (60.6%). In an assembly stress test, AIVA detected all 644 corrupted genome-build records and self-corrected them, resulting in 68% of these mislabeled variants being correctly classified. Benchmarking prioritization across 1,396 simulated rare disease cases, AIVA ranked the causal gene first in 66.7% of cases, versus 59.5% for LIRICAL and 28.4% for Exomiser, by leveraging its real-time literature review and agentic capabilities.

Conclusions

To our knowledge, AIVA is among the first agentic platforms to perform both variant prioritization and ACMG/AMP classification, outperforming established tools on both tasks. By retrieving functional data, VCEP specifications, and phenotype evidence at analysis time, it applies the full set of ACMG/AMP criteria, including the literature-dependent ones that rule-based classifiers cannot evaluate. AIVA can identify and self-correct errors in the input like assembly mismatch, keeping results consistent. In one conversational, literature-grounded workflow, AIVA closes the gap between automated classifiers and expert tertiary review. AIVA is available at https://chat.aivaportal.com.

Background

Rare diseases affect an estimated 300 million people worldwide, but diagnosis remains slow. Patients experience a diagnostic odyssey averaging 5 to 7 years, consulting up to eight physicians and receiving two to three misdiagnoses before a molecular cause is identified [1, 2]. Genomic sequencing has

reduced the cost and turnaround of data generation, shifting the bottleneck from sequencing to interpretation. Secondary analysis has largely improved the accuracy of variant calling. Deep learning-based callers such as Google's DeepVariant and DeepSomatic, along with Illumina's DRAGEN, achieve near-equivalent accuracy for germline SNVs and indels, with pangenome-aware modes reducing errors by up to 25% [3, 4, 5]. A whole-exome run now produces 100,000 to 200,000 variants per patient, and whole-genome approximately 5 million [6]. The challenge is identifying the pathogenic variant(s) among thousands of variants per patient that are associated with their phenotype.

Clinical interpretation is highly labor-intensive and distributed among PhD-level analysts and genetic counselors. A well-characterized variant may take a few minutes to evaluate, while a poorly characterized variant with conflicting database entries can require anywhere from 30 minutes to several hours of manual review, synthesizing publications and querying databases like ClinVar, OMIM, and gnomAD [7]. Accordingly, 72.0% of clinical geneticists cite lack of time as a primary barrier to thorough interpretation [8]. Several computational tools were developed to assist in both variant classification and prioritization. Rule-based classifiers such as InterVar [9] and BIAS-2015 [10] automate the ACMG/AMP guidelines [11] by integrating population frequency, conservation scores, and in silico predictors. VarSome aggregates over 150 databases to automate approximately 82% of ACMG criteria [12], while Franklin incorporates AI-driven phenotype weighting [13]. On the other hand, phenotype-driven gene prioritization tools such as LIRICAL [14], Exomiser [15], and AMELIE [16] rank candidate genes against curated knowledge graphs from OMIM, Orphanet, and the Human Phenotype Ontology [17], with AMELIE also applying natural language processing to the literature. Commercial platforms, including Illumina's Emedgene and Fabric GEM, report 2- to 5-fold throughput improvements and up to 97% accuracy in trio analyses [18, 19]. A common limitation persists across these tools: they filter and rank candidate variants and genes (~ 5M to ~ 50), but the interpretive review of these filtered variants still occurs outside the platforms. Analysts return to PubMed, UniProt, HGMD, etc., to evaluate the evidence for variant-to-phenotype associations and build the case for or against pathogenicity. Moreover, these automated classifiers apply generic ACMG/AMP rules, or gene-specific rules only where these have been hard-coded for a particular gene or disease area such as inherited cardiac conditions [20]. Identifying and applying the appropriate ClinGen Variant Curation Expert Panel (VCEP) specification for any gene/disease model when applicable is crucial for accurate variant classification and reporting. Most platforms also confine users to a fixed interface, displaying only the annotations generated by their internal pipeline, with no way to integrate a laboratory's own annotations.

Large language models (LLMs) take a different approach. They encode knowledge from biomedical corpora that analysts manually consult and can synthesize information across gene function, protein structure, disease mechanisms, and phenotype ontologies in a way that mirrors the reasoning of a clinical geneticist. Their conversational interface makes them suited to serve as a second opinion: an analyst can describe a case in plain language and receive a reasoned assessment. GPT-based models with retrieval-augmented generation have performed variant annotation with reasonable accuracy [21], and specialized models such as GP-GPT have shown promise in gene-phenotype mapping [22]. Standalone LLMs, however, are prone to hallucination. Without grounding in verified data sources, an

LLM may generate fabricated variant-disease associations, cite nonexistent publications, or apply ACMG criteria to confabulated evidence.

Agentic AI systems, in which LLMs are augmented with structured tool access, can provide grounded responses. VarChat [23] pairs an LLM with a curated literature retrieval engine to generate evidence-backed variant summaries; however, it only operates on individual variants or phenotype descriptions rather than processing complete VCF files for end-to-end case-level interpretation. DeepRare [24] integrates over 40 specialized tools to produce ranked diagnostic hypotheses with traceable reasoning chains. However, it requires local deployment with large external databases (gnomAD, CADD, etc.) and doesn't support customizable annotation environments that allow laboratories to retain their own pipeline outputs alongside the system's analysis.

We developed AIVA (AI-powered Variant Analysis), an end-to-end agentic platform designed to address the interpretative gap through a conversational, literature-grounded interface. In contrast to static tools, AIVA performs automated, evidence-based analysis by querying biomedical knowledge bases and reviewing the literature to generate reasoned assessments, accompanied by citations, for every claim. By supporting custom annotation imports and providing a scalable 'second opinion' that is resilient to review fatigue, AIVA enables clinical laboratories to resolve the tertiary analysis bottleneck while maintaining the rigorous traceability required for molecular diagnosis.

Implementation

1. AIVA: AI-powered Variant Analysis

AIVA is an explainable tertiary analysis agentic system that uses state-of-the-art LLMs to identify causal variants from exome/genome sequencing data based on clinical notes. The following subsections describe the platform architecture and features for scalable and explainable genetic disease interpretation.

1.1 Data Input and Annotation Environment

AIVA supports a flexible input environment that does not constrain users to a single pipeline. Users can begin with either FASTQ (raw sequencing output) files or pre-called Variant Call Format (VCF) files from targeted panels, whole-exome, or whole-genome sequencing. AIVA uses NVIDIA Parabricks [25] under the hood for the secondary analysis pipeline to process FASTQ to VCFs. AIVA also supports directly uploading pre-annotated VCFs generated by any standard pipeline (e.g., nf-core/sarek) or by any annotation tool (e.g., VEP, SnpEff). This is to support users bringing any in-house annotations on top of AIVA annotations for variant interpretation.

The benchmark dataset comprised variants curated from the ClinGen Evidence Repository (eRepo), spanning multiple disease categories and expert panels. Table 2 summarizes the composition of this dataset by disease category, number of genes, VCEPs, and pathogenicity class distribution.

1.2 Conversational Interface and Agentic Architecture

Users interact with AIVA through a conversational interface in which they query variants from samples (singleton, trios, cohorts), validate against BAM files, flag, and classify them according to ACMG/AMP criteria. Users can also request supporting rationales in natural language, describing the agent's analytical logic, including filtering criteria, annotation preferences, and decision frameworks. For each query, the agent formulates a reasoning plan, retrieves relevant evidence from specialized genomics tools in real time, and synthesizes a response grounded in the retrieved data to accurately answer the user's query. Beyond variant-level records, the agent can query the underlying BAM alignments directly, inspecting read depth, allele balance, strand bias, and mapping quality, etc. It can also confirm de novo status in trios, so read-level evidence can be interrogated conversationally within the same session as the VCF rather than manual review for each variant. The same alignment access supports validation of copy-number variants (CNVs), structural variants (SVs), read-backed phasing, and multi-nucleotide variants (MNVs).

1.3 Extensibility and Interoperability

AIVA is also designed to be easily integrated within existing clinical genomics workflows. (i) AIVA offers API access, so users can easily upload FASTQs or VCFs from their secondary analysis pipelines for downstream analysis. (ii) Users can also extend AIVA's reasoning capabilities by connecting external Model Context Protocol (MCP) servers or proprietary knowledge bases. For example, internal variant databases, in-house functional assay repositories, or laboratory information systems can be connected via MCP servers so that AIVA can use these toolsets during an analysis session. (iii) AIVA can be added as an MCP tool in other agentic platforms like Claude or Codex, exposing its variant interpretation capabilities as a callable service for analysis.

2. Benchmarking

We benchmarked AIVA's performance on both prioritization and Classification workflows. The complete study design for both benchmarks is illustrated in Fig. 2.

2.1 Phenotype-Aware Variant prioritization

2.1.1 Synthetic Patient Cohort

We extracted a cohort of 1,396 UDN-derived simulated rare disease cases using the framework described by Alsentzer et al. [26, 27]. Each case consists of a VCF file (standardized to GRCh37 (hg19)) with a causal variant, accompanied by a set of Human Phenotype Ontology (HPO) terms [17]. To simulate real-world diagnostic complexity, phenotype profiles included core diagnostic HPO terms, supporting terms, and deliberate noise terms (distractors or imprecise clinical descriptions). Each case provided a set of candidate genes (a mean of 14, SD 3.5) generated by the simulation framework: the

causal gene together with deliberately seeded distractor genes, such as phenotypically similar disease genes and previously reported pathogenic but phenotype-irrelevant genes. Each candidate gene was represented by a single high-impact SNV. Because the candidate list is dominated by plausible distractors, ranking the causal gene is deliberately hard; this identical candidate set was provided, without further filtering, to all three tools.

2.1.2 Tool Implementation and Configurations

We benchmarked AIVA against two open-source, well-maintained prioritization algorithms, LIRICAL and Exomiser. LIRICAL (v2.2.1) was run in global mode using the GRCh37 (hg19) 2406 data release, enabling unbiased ranking across the entire Online Mendelian Inheritance in Man (OMIM)/Orphanet knowledge base; genes were ranked by post-test Probability. Exomiser (v14.0.0) was configured with the hiPhivePrioritiser (human+mouse+fish models) and omimPrioritiser, with a frequency filter of 2.0% applied; candidates were ranked by the EXOMISER_GENE_COMBINED_SCORE. AIVA used Google Gemini 3.0 Pro Preview as a default model for this benchmark. Rankings were generated in a single inference pass at temperature 0.3. (see Supplementary Methods S1 for the complete prompt template)

2.1.3 Performance Metrics

Recall@K (K = 1, 5, 10, 20) serves as the primary metric, defined as the number of cases where the causal gene appears within the top K-ranked outputs (from variant-level inputs). We also assessed diagnostic accuracy across different disease types using the categories from Alsentzer et al., ranging from known to new disease-gene associations. In addition, for cases that AIVA failed to rank in top-1, we characterized the composition of the gene candidates by reporting the distractor-gene type distribution at top-1 and across disease categories.

2.2. Variant Classification Benchmark

2.2.1 Benchmark Dataset

We used ClinGen (eRepo)- classified variants as our benchmark dataset, derived from Eisenhart, J., Brickey, W., et al. [10], which aggregates variant classifications from Variant Curation Expert Panels (VCEPs) following ACMG/AMP 2015 guidelines [11]. The extracted cohort included 8,387 variants across 108 genes and 11 disease categories, curated by 35 VCEPs. All genomic coordinates in the BIAS-2015 eRepo dataset were already provided in GRCh37 (hg19), and all tools in this benchmark were run against GRCh37 coordinates. We collapsed the ACMG classes into three directional classes for comparison: Pathogenic/Likely Pathogenic (P/LP; n = 3,549, 42.3%), Variant of Uncertain Significance (VUS; n = 2,779, 33.14%), and Benign/Likely Benign (B/LB; n = 2,059, 24.56%).

Table 2

Variant classification benchmark dataset composition. P/LP = Pathogenic/Likely Pathogenic; B/LB = Benign/Likely Benign; VUS = Variant of Uncertain Significance. Percentages indicate class distribution within each category.

Disease category	# variants	# VCEPs	Unique genes	Pathogenic	VUS	Benign
Cancer Predisposition	2,382	9	17	554 (23.3%)	997 (41.9%)	831 (34.9%)
Metabolic Disorders	2,677	7	16	1,634 (61.0%)	804 (30.0%)	239 (8.9%)
Cardiovascular Disorders	341	3	3	161 (47.2%)	101 (29.6%)	79 (23.2%)
Hematologic Disorders	695	4	6	383 (55.1%)	176 (25.3%)	136 (19.6%)
Immunodeficiency Disorders	379	1	8	144 (38.0%)	150 (39.6%)	85 (22.4%)
Neurodevelopmental Disorders	387	1	6	98 (25.3%)	50 (12.9%)	239 (61.8%)
Neuromuscular Disorders	524	3	12	191 (36.5%)	233 (44.5%)	100 (19.1%)
Ophthalmologic Disorders	332	2	2	119 (35.8%)	157 (47.3%)	56 (16.9%)
Hearing Disorders	196	1	11	94 (48.0%)	45 (23.0%)	57 (29.1%)
RASopathies	365	1	17	126 (34.5%)	42 (11.5%)	197 (54.0%)
Other	109	3	10	45 (41.3%)	24 (22.0%)	40 (36.7%)
TOTAL	8,387	35	108	3,549 (42.3%)	2,779 (33.1%)	2,059 (24.5%)

2.2.2 Tool Implementation and Configurations

We compared AIVA against two open-source and well-maintained classification tools representing different methodological approaches. AIVA is an agentic system powered by the Google Gemini 3.0 Pro Preview model, queried via a REST API, providing 3-class classifications with supporting rationales. Citations in AIVA's output are programmatically attached from the underlying tool results (the retrieved text and its source) rather than generated by the language model, so the model surfaces evidence rather than hallucinating references. All AIVA classifications were produced in a single inference pass at temperature 0.3. BIAS-2015 v2.1.1 [10] (Rule-based) is a standardized ACMG/AMP classifier that automates 19 criteria based on computational evidence, including population frequency (gnomAD), conservation (PhyloP, GERP), and computational pathogenicity predictors (CADD, REVEL) [28], while relying on manual user input to incorporate disease-specific biological context or phenotypic criteria. InterVar [9] (Rule-based) is another open-source automated classifier accessed via the wInterVar API that integrates databases such as ClinVar and dbSNP with computational evidence. All tools were run using default parameters to ensure a standardized benchmark against the 8,387 variants. All three tools received identical inputs (variant coordinates, target gene, disease/phenotype context, and mode of inheritance).

2.2.3 Performance Evaluation

Metrics were computed using scikit-learn to assess overall tool performance and across disease categories; macro F1 score was chosen as the primary metric to ensure balanced weighting across the three directional classes. Per-category macro F1 scores were computed independently for each disease category using all variants within that category, with equal class weighting across the three directional classes (P/LP, VUS, B/LB). Class-specific precision, recall (sensitivity), specificity, and F1 scores were calculated for each individual clinical classification class (i.e. P/LP, VUS, B/LB). Mis-classification analysis utilized Sankey flow diagrams (Figure S1) to identify systematic misclassification patterns, such as the tendency to upgrade benign variants to VUS.

Results

We benchmarked AIVA on two levels - variant prioritization and classification. This benchmark was designed to mimic real-world variant analysis by ranking the causal gene within a provided candidate set for prioritization and then accurately classifying them based on the patient's phenotype, inheritance, zygosity, etc.

1. Variant prioritization

We evaluated AIVA on identifying the causal gene (from variant-level inputs) in 1,396 UDN-simulated rare disease cases against well-maintained and open-source tools like Exomiser and LIRICAL.

1.1 Overall Performance

Figure 3A summarizes key primary metrics (Recall@K for K = 1, 5, 10, 20), with K being the position of the causal gene in each tool's ranked output. AIVA ranked the true causal gene first in 66.7% of cases (Top-1 accuracy), compared with 59.5% for LIRICAL and 28.4% for Exomiser across 1,396 patients. Bold cells indicate the best-performing tool within each row. Lower performance of these tools is attributed to the databases enriched, where phenotype-based tools operating from known disease databases are known to underperform [14, 15]. At top-20, AIVA solved 94.4% of cases compared to 86.4% by LIRICAL and 90.0% by Exomiser. We also compared the mean and median rank of causal genes by each tool. LIRICAL placed it at median rank 1 (mean rank 2.0, range 1–18), while Exomiser placed the true causal gene at median rank 6 (mean rank 6.5, range 1–28), and AIVA's median rank was 1 (mean rank 2.0, range 1–20).

In addition, we compared whether tools failed to return the causal gene in the ranked list, which we reported as the causal gene absent rate. AIVA's causal gene absent rate of 5.6% was the lowest of the three tools, compared with 13.6% for LIRICAL and 9.4% for Exomiser.

1.2 Per-Category Performance

We extended the benchmarking to compare tools against different disease categories. Figure 3A highlights Top-20 accuracy for all three tools across different categories from already known gene-disease association to novel gene-disease associations. AIVA leads in all five categories: Known Gene Disease (96.4%), Known Gene + Known Disease (86.3%), Known Gene + New Disease (94.7%), New Gene + Known Disease (99.2%), and New Gene + New Disease (93.6%, tied with LIRICAL at 93.6%).

Exomiser's performance declines sharply in new-gene categories: 20.7% Top-1 for New Gene + Known Disease and 8.8% for New Gene + New Disease, compared with 38.9% in Known Gene Disease cases. This likely reflects their dependence on curated knowledge bases with limited coverage of newly described associations, whereas AIVA can perform a literature review in real time to identify new gene-disease associations.

1.3 Distractor-Gene Analysis

To understand the nature of AIVA's rank-1 predictions that are not causal variants, we classified the displacing gene/variant in each of the 387 non-rank-1 cases by distractor type (Fig. 3B). We observed pathogenic-but-phenotype-irrelevant variants (33.2%, n = 132), i.e., ClinVar classified pathogenic variants irrespective of gene-phenotype association were ranked above the causal variant. The second most common type was phenotype distractors (26.7%, n = 106), i.e., genes that are associated with the distractor phenotype terms. Together, these two categories account for nearly 60% of all missed rank-1 predictions.

We extended this approach for each disease category to observe how and why AIVA ranks distractor variants above causal variants. Figure 3C shows that pathogenicity and phenotype distractors are the main contributors in all disease categories. In Known-Gene + New-Disease cases, phenotype distractors account for 42% of all mispredictions (34 out of 81), the highest proportion across all groups and significantly above the global average of 26.7%. This pattern reflects the inherent challenge of

distinguishing phenotypically overlapping conditions when the underlying gene–disease relationship is novel. In comparison, pathogenic but phenotype-irrelevant genes dominate the predictions in Known-Gene-Disease (39%), New-Gene + Known-Disease (33%), and New-Gene + New-Disease (33%) groups. These results suggest that AIVA's variant pathogenicity reasoning can, at times, outweigh its phenotypic inference when a competing gene is a previously reported pathogenic variant. The remaining non-rank-1 predictions across all categories were distributed, such as unclassified background variants ("not labelled", 17.6%) and universal distractors (highly plausible genes that are not clinically relevant, 8.6%).

One example of how phenotype distractors affect AIVA's performance is illustrated by a case of undetermined early-onset epileptic encephalopathy. The patient presented with hypsarrhythmia, seizures, intellectual disability, developmental regression, abnormal corpus callosum morphology, Peters anomaly, optic atrophy, and hypodontia. The true causal gene was GABRB2, encoding the GABA-A receptor β2 subunit; however, AIVA ranked POGZ (score 0.95) on top in the resulting ranked list over GABRB2 (score 0.75). The reasoning is that the phenotypes, namely anterior segment dysgenesis, corpus callosum abnormality, dental anomalies, and congenital heart defect, were more consistent with POGZ-associated White-Sutton syndrome than with an isolated GABA receptor channelopathy. This example illustrates how phenotypic overlap can lead AIVA to rank a plausible alternative gene above the known causal gene.

2. Variant Classification

Dataset composition and full breakdown by category are summarized in Table 2 in the Implementation section. The dataset comprised 8,387 variants spanning 108 genes across 11 disease categories, curated by 35 VCEPs.

2.1 Classification Performance

Figure 4 presents the full performance summary by class and disease category. AIVA achieved an overall macro F1 of 80.5% across all 8,387 variants, outperforming both BIAS-2015 v2.1.1 (75.3%) and InterVar (60.6%). We also compared per-class performance for AIVA, with Pathogenic/LP F1 = 87.4%, Benign/LB F1 = 80.9%, and VUS F1 = 73.3%, outperforming other classifiers. Figure S1A shows AIVA's classification behavior, as we observed that AIVA often identifies some evidence to reclassify variants mainly because it has real-time access to literature search. Examining the distribution by true class (Figure S1B), of true Pathogenic/LP variants, 88.5% were correctly classified; 11.5% were downgraded to VUS, and fewer than 0.1% (n = 1) were misclassified as Benign/LB. Of true Benign/LB variants, 72.3% were correctly classified as benign; 26.8% were upgraded to VUS, reflecting AIVA's capability to identify evidence linking the variant/gene to phenotype. VUS classification was stable at 78.9% concordance, with 16.3% upgraded to P/LP and 4.8% downgraded to B/LB. The proportional breakdown of reclassification modes (Figure S1C) confirms that conservative adjacent-category shifts dominate (Benign to VUS: 34.5%, VUS to P/LP: 29.7%), while binary clinical discordances (Pathogenic↔Benign) remain < 1.3%.

This phenotype- and gene-aware behavior is illustrated by PPP1CB c.166G > C (p.Ala56Pro), which AIVA classified as Likely Pathogenic by retrieving and applying the gene-specific ClinGen RASopathy VCEP specification (GN128 v1.3.0) [29] rather than generic ACMG thresholds (Box 1). It awarded PS2 at Strong from a confirmed de novo trio in the literature, with PM2, PS4, and PP2 at Supporting, while suppressing PVS1, PS3, PM1, and PP4 (all marked not applicable for this gain-of-function gene) and withholding PP3 because the VCEP requires REVEL of at least 0.7, here 0.463, even though CADD (26.2) exceeds the generic threshold. A deterministic classifier on generic thresholds, without the de novo evidence, would not reproduce this call. Because the VCEP specification and the supporting de novo report were retrieved and cited from tool outputs rather than recalled by the model, this example demonstrates that AIVA's calls rest on evidence surfaced at analysis time rather than on the language model's parametric knowledge.

Box 1. ACMG/AMP criteria with evidence applied to PPP1CB c.166G > C (p.Ala56Pro) under the ClinGen RASopathy VCEP (GN128 v1.3.0). Final classification: one Strong plus three Supporting criteria yields Likely Pathogenic (ACMG points + 7).

Criterion	Outcome	Points	Evidence
PS2	Applied (Strong)	+ 4	Confirmed de novo with maternity and paternity verified; trio exome sequencing (Gripp et al. 2016) [30]
PS4	Applied (Supporting)	+ 1	Occurrence in affected RASopathy cases [30, 31, 32]
PM2	Applied (Supporting)	+ 1	Absent from gnomAD v4 (genome and exome)
PP2	Applied (Supporting)	+ 1	Missense constraint, missense Z = 6.71 (VCEP threshold > 3.09)
PVS1	Not applicable	—	Loss of function is not the disease mechanism; PPP1CB acts by gain of function per VCEP
PS3	Not applicable	—	No approved functional assay for PPP1CB per VCEP
PM1	Not applicable	—	Not applicable for PPP1CB per VCEP
PP3	Not met	—	REVEL 0.463 below the VCEP threshold of 0.7; generic CADD 26.2 not used
PP4	Not applicable	—	Replaced by PS4 in the VCEP specification

We compared the performance of all 3 classifiers by disease category to observe if there is any bias toward certain disease categories. AIVA outperformed both BIAS-2015 and InterVar in 9 of 11 disease categories (Fig. 4A). The largest advantages calculated in percentage points (pp) were in Hematologic (+ 13.1 pp over BIAS-2015), Cardiovascular (+ 8.8 pp), Neuromuscular (+ 8.5 pp), and Other (+ 10.7 pp). AIVA trailed BIAS-2015 by a small margin in Hearing Loss (− 1.0 pp) and Neurodevelopmental disorders (− 1.9 pp). A detailed investigation into the variants in these categories revealed that most are classified

as benign, with Neurodevelopmental Disorders (61.8%) and Hearing Loss (29.1%) being the most common. This is likely because AIVA often reclassifies benign variants as VUS (20% as shown in Figure S1A) through real-time literature review.

2.2 Assembly Self-Correction Stress Test

To evaluate AIVA's resilience to real-world input errors, we deliberately annotated 644 of the 8,387 variants (7.7%) by tagging the wrong genome builds. The benchmark reference remained GRCh37 (hg19), but for this stress test we paired GRCh37-labeled records with GRCh38 (hg38) positions, creating a coordinate mismatch that could cause incorrect genomic context lookups. The test was designed to probe whether AIVA could detect and recover from malformed inputs without external intervention. The other two tools, BIAS-2015 and InterVar, were only evaluated on correct build coordinates because they are deterministic tools with no coordinate resolution capabilities. Evidence logs from these records confirmed detection, with AIVA flagging "CRITICAL ASSEMBLY ERROR DETECTED" in its reasoning for each affected variant.

AIVA detected all 644 discrepancy records (100.0%) and applied a three-step self-correction workflow: identify an assembly conflict, resolve coordinates using the most likely reference assembly, and then reclassify using the corrected coordinates. Of these 644 records, 438 (68.0%) were resolved (matching ClinGen ground truth) and classified correctly according to ACMG criteria, while 206 (32.0%) were resolved but still misclassified (Fig. 4B). The largest and most difficult subgroup was RUNX1 (51.4% of discrepant records) because different builds represent different transcripts/genes. In this subgroup, 71.6% of corrupted records were classified correctly after self-correction, compared with 71.5% of uncorrupted RUNX1 records.

Discussion

Our benchmark deliberately targets rare and novel disease cases, the cohort spans from well-characterized to previously unknown gene–disease associated variants. This is where phenotype-driven tools that depend on curated knowledge graphs are weakest, and our results back the claim. AIVA ranked the causal gene first in 66.7% of cases and failed to return it in only 5.6%, the lowest miss rate of the three tools. LIRICAL was competitive when the causal gene was already well described in OMIM or Orphanet, but Exomiser dropped to 28.4% top-1 overall and just 8.8% on new gene/new disease cases, because it relies on existing phenotype–gene associations that are sparse or absent for rare conditions. AIVA does not rely on those priors, so it is best used alongside these tools rather than as a replacement: the bounded tools stay strong when a clear prior disease hypothesis exists, and AIVA adds coverage when it does not.

In classification, AIVA reached a macro F1 of 80.5%, ahead of BIAS-2015 (75.3%) and InterVar (60.6%). Its advantage was largest on criteria that need literature and clinical context, PS3/BS3 (functional evidence), PM3/BP2 (phase), and PP1/BS4 (segregation), which rule-based tools cannot apply without curated, gene-specific rule sets. Unlike aggregators such as VarSome or Franklin, which combine many

databases into a single automated call, AIVA reasons over the retrieved evidence and shows its work for each criterion. Much of the remaining disagreement with the VCEP labels was not error but scope: AIVA applied criteria that the deterministic tools leave out, or it lacked the private evidence the expert panels used. This makes AIVA a decision-support layer that flags variants worth a second look and leaves the final call to the analyst.

A distinct capability underlies this advantage: AIVA retrieves and applies the appropriate gene-specific ClinGen VCEP specification at the time of analysis rather than generic ACMG thresholds. In the PPP1CB example (Box 1), it applied the RASopathy VCEP (GN128 v1.3.0), awarding PS2 from a confirmed de novo trio, withholding PP3 because that VCEP requires REVEL of at least 0.7 (here 0.463), and suppressing PVS1 for this gain-of-function gene. Existing automated classifiers apply VCEP-specific rules only where they have been hard-coded for a particular gene or disease area, for example CardioClassifier for inherited cardiac conditions [20], whereas AIVA applies the correct specification across arbitrary genes and panels without per-gene engineering.

A main reason for these gains is real-time literature retrieval. HPO, OMIM, and PanelApp are updated periodically and can lag the primary literature by months to years, whereas AIVA queries PubMed now of interpretation. This mattered most in novel gene–disease cases, where curated databases assign the causal gene a low prior, but recent papers already support the causal gene.

AIVA also sits at a different point in the workflow than other agentic and LLM-based tools. VarChat [23] generates evidence-backed summaries for a single variant or phenotype query but does not process a full VCF or rank candidates across a case. DeepRare and SHEPHERD [24, 33] rank candidate diseases or genes but do not classify individual variants against ACMG criteria, and AutoPM3 [34] retrieves evidence for a single criterion (PM3) rather than running a case end-to-end. AIVA does both in one workflow, from VCF to a step-by-step ACMG rationale, and its assembly self-correction (reconciling GRCh37 and GRCh38 coordinates) handles a failure mode other LLM-based tools ignore, keeping results consistent across pipelines.

Limitations

Several limitations should be noted. First, the prioritization benchmark evaluates ranking within a provided candidate set of, on average, 14 genes rather than de novo filtering from a complete WES/WGS VCF. This is the limitation of the simulated data that we obtained but it mimics the standard filters applied while reviewing a rare disease case. However, AIVA can apply the same filters from a whole VCF. Second, all classifications were generated in a single inference pass of Google Gemini 3.0 Pro at temperature 0.3; run-to-run variability was not characterized as the model is now deprecated, and reported differences are not accompanied by confidence intervals or paired significance testing. Finally, AIVA is a decision-support tool intended to assist, not replace, expert review; it has not undergone prospective clinical validation or regulatory evaluation, and the classification labels reflect benchmark concordance rather than established clinical performance.

Conclusions

To our knowledge, AIVA is among the first end-to-end agentic platforms to perform both variant prioritization and ACMG classification, and it outperformed established tools on both tasks: top-1 prioritization of 66.7% (versus 59.5% and 28.4%) and a classification macro F1 of 80.5% (versus 75.3% and 60.6%). It applies the full set of ACMG/AMP criteria, including the literature-dependent ones (PS3, PM3, PP1) that rule-based classifiers cannot evaluate, by retrieving functional data, expert-panel assertions (VCEPs), and phenotype evidence at the time of analysis. Its assembly self-correction keeps results consistent across genome builds (GRCh37 and GRCh38), a failure mode that other tools do not address. By combining real-time literature retrieval, phenotype-driven prioritization, and step-by-step ACMG rationale in one conversational workflow, AIVA closes the gap between automated classifiers and expert tertiary review and supports scalable evidence-based tertiary analysis for rare disease interpretation.

Abbreviations

SOTA
State-of-the-art
ACMG
American College of Medical Genetics and Genomics
AMP
Association for Molecular Pathology
VCF
Variant Call Format
HPO
Human Phenotype Ontology
LLM
Large Language Model
VCEP
Variant Curation Expert Panel
VUS
Variant of Uncertain Significance
P/LP
Pathogenic/Likely Pathogenic
B/LB
Benign/Likely Benign
MRR
Mean Reciprocal Rank
UDN
Undiagnosed Diseases Network
RAG

Retrieval-Augmented Generation
MCP
Model Context Protocol

Declarations

Availability and requirements

Project name: AIVA

Project home page: https://chat.aivaportal.com

Operating system(s): Platform independent (web-based)

Other requirements: Institutional access credentials

License: Proprietary (Mamidi Health)

Any restrictions to use by non-academics: Commercial use requires a licence agreement with Mamidi Health

Availability of data and materials

The benchmark datasets used in this study are publicly available. The ClinGen variant dataset (n = 8,387) was obtained from the benchmark dataset used by Eisenhart et al. (2025) (BIAS-2015 v2.1.1; Genome Medicine; DOI: 10.1186/s13073-025-01581-y). The simulated patient cohort (n = 1,396) was derived from the Benchmark 2 dataset published by Alsentzer et al. (2023), available online (https://zenodo.org/records/8190872, https://doi.org/10.5281/zenodo.8190872, accessed 10 Jan 2026). AIVA is accessible via its web portal at https://chat.aivaportal.com. The benchmarking analysis codebase for variant classification (https://github.com/MHSPL/AIVA-var-bench) and variant prioritization (https://github.com/MHSPL/AIVA-rank-bench) are publicly available.

Ethics approval and consent to participate

Not applicable. This study used publicly available benchmark datasets and did not involve human participants directly.

Consent for publication

Not applicable.

Competing interests

Tarun Mamidi is the founder of Mamidi Health. Arun K. Boddapati declares no competing interests.

Funding

This work received no external funding. The AIVA platform was developed independently by the authors through Mamidi Health.

Authors' contributions

TM conceived the study, designed the AIVA platform architecture and platform development. AKB contributed to designing and implementing benchmarking analysis and manuscript writing. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the broader rare disease and clinical genomics community for motivating this work. We are grateful to the developers and maintainers of the open-source tools benchmarked in this study, including BIAS-2015, InterVar, LIRICAL and Exomiser, whose contributions to the field enabled meaningful comparative evaluation. We also acknowledge the patients and families affected by rare genetic diseases, whose diagnostic challenges inspired the development of AIVA.

References

1. Genomics Education Programme. The diagnostic odyssey in rare disease. Health Education England. Available from: https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/the-diagnostic-odyssey-in-rare-disease/
2. EveryLife Foundation for Rare Diseases. The cost of delayed diagnosis in rare disease. 2023. Available from: https://everylifefoundation.org/wp-content/uploads/2023/09/EveryLife-Cost-of-Delayed-Diagnosis-in-Rare-Disease_Final-Full-Study-Report_0914223.pdf
3. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7.
4. Accurate somatic small. variant discovery for multiple sequencing technologies with DeepSomatic. Nat Biotechnol. 2025. 10.1038/s41587-025-02839-x.
5. Comprehensive genome analysis. and variant detection at scale using DRAGEN. Nat Biotechnol. 2024. 10.1038/s41587-024-02382-1.
6. Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. NPJ Genom Med. 2020;5:47.
7. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.
8. Wain KE, Azzariti DR, Goldstein JL, et al. Variant interpretation is a component of clinical practice among genetic counselors in multiple specialties. Genet Med. 2020;22(4):785–92.

9. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100(2):267–80.
10. Eisenhart C, Brickey R, Nadon B, et al. Automating ACMG variant classifications with BIAS-2015 v2.1.1: algorithm analysis and benchmark against the FDA-approved eRepo dataset. Genome Med. 2025;17:148. 10.1186/s13073-025-01581-y.
11. 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.
12. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80.
13. Genoox. Franklin: community-driven variant interpretation platform. Available from: https://franklin.genoox.com
14. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Banka S, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Am J Hum Genet. 2020;107(3):403–17.
15. Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet. 2016;99(3):595–606.
16. Birgmeier J, Haeussler M, Deisseroth CA, Steinberg EH, Baca AR, Jagadeesh KA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020;12(544):eaau9113.
17. Kohler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207–17.
18. Illumina. Emedgene: germline variant interpretation. Available from: https://www.illumina.com/products/by-type/informatics-products/emedgene.html
19. De La Vega FM, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome Med. 2021;13:153.
20. Whiffin N, Walsh R, Govind R, Edwards M, Ahmad M, Zhang X, et al. CardioClassifier: disease- and gene-specific computational decision support for clinical genome interpretation. Genet Med. 2018;20(10):1246–54.
21. Lu S, Cosgun E. Boosting GPT models for genomics analysis: generating trusted genetic variant annotations and interpretations through RAG and fine-tuning. Bioinform Adv. 2025;5(1):vbaf019.
22. Lyu Y et al. GP-GPT: large language model for gene-phenotype mapping. arXiv:2409.09825. 2024.
23. De Paoli F, Berardelli S, Limongelli I, Rizzo E, Zucca S. VarChat: the generative AI assistant for the interpretation of human genomic variations. Bioinformatics. 2024;40(4):btae183.

24. Zhao W, Wu C, Fan Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature. 2026. 10.1038/s41586-025-10097-9.
25. NVIDIA. Improve variant calling accuracy with NVIDIA Parabricks. NVIDIA Technical Blog. 2025. Available from: https://developer.nvidia.com/blog/improve-variant-calling-accuracy-with-nvidia-parabricks/
26. Alsentzer E, Finlayson SG, Li MM, Kobren SN, Kohane IS. Simulation of undiagnosed patients with novel genetic conditions. Nat Commun. 2023;14:6403.
27. Alsentzer E, Finlayson S, Li M, Kobren S, Kohane I. rare-disease-simulation: simulate patients with rare genetic conditions. Zenodo. 2023. 10.5281/zenodo.8190872. Available from: https://zenodo.org/records/8190872
28. 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.
29. ClinGen RASopathy Variant Curation Expert Panel. ACMG/AMP criteria specifications for the RASopathy genes (affiliation GN128), v1.3.0. ClinGen Criteria Specification Registry. Available from: https://cspec.genome.network/
30. Gripp KW, Aldinger KA, Bennett JT, et al. A novel rasopathy caused by recurrent de novo missense mutations in PPP1CB closely resembles Noonan syndrome with loose anagen hair. Am J Med Genet A. 2016;170(9):2237–47.
31. Epileptic. spasms in PPP1CB-associated Noonan-like syndrome: a case report with clinical and therapeutic implications. BMC Neurol. 2018;18:150. PMID:30236064.
32. Case report. identification and clinical phenotypic analysis of a novel mutation of the PPP1CB gene in NSLH2 syndrome. Front Behav Neurosci. 2022;16:987259. PMID:36160684.
33. Alsentzer E, Li MM, Kobren SN, Noori A, Undiagnosed Diseases Network, Kohane IS, et al. Few-shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. NPJ Digit Med. 2025;8(1):380.
34. Li S, Wang Y, Liu CM, Huang Y, Lam TW, Luo R. AutoPM3: enhancing variant interpretation via LLM-driven PM3 evidence extraction from scientific literature. Bioinformatics. 2025;41(7):btaf382.

Figures

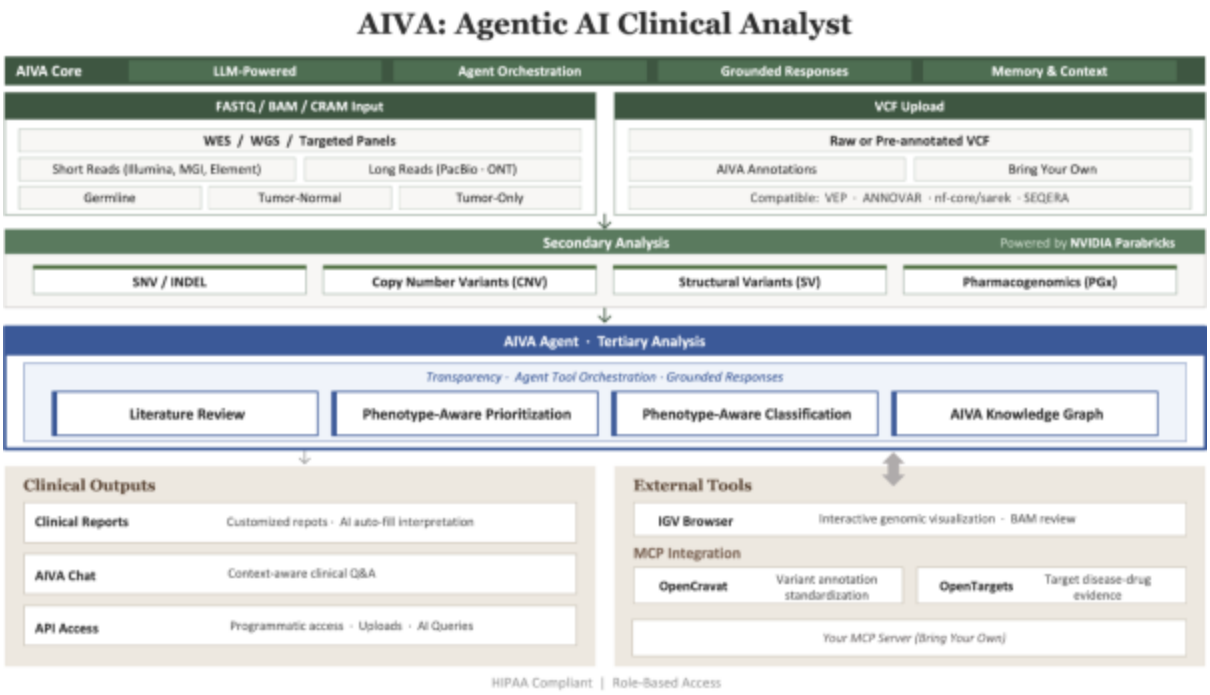

Figure 1

AIVA system architecture. The AIVA platform integrates four core components: (i) Customizable input with either FASTQ or VCFs (pre-annotated or raw), (ii) GPU-accelerated secondary pipeline using NVIDIA Parabricks (iii) Annotations and agentic analysis orchestrated by tools accessing clinical databases, and (iv) Clinical outputs include customized reports, chat-based reasoning, and API access. AIVA is designed to support HIPAA compliance, operating under business associate agreements with zero-data-retention terms across all LLM providers. The agent can also query the underlying BAM files directly to examine read-level evidence within the same session as the VCF.

Figure 2

Benchmark study design overview. (A) Variant prioritization workflow. A cohort of 1,396 simulated rare disease patients was used to benchmark LIRICAL (v2.2.1), Exomiser (v14.0.0), and AIVA. Each tool received HPO phenotype terms and an annotated VCF (GRCh37). Ranked gene outputs were evaluated by Recall@K (ranked position, K = 1, 5, 10, 20). **(B) Variant Classification workflow**. ClinGen Evidence

Repository (eRepo; 8,387 variants, 108 genes, 35 VCEPs, 11 disease categories) served as ground truth. Each variant record (comprising GRCh37 coordinates and gene/disease context) was independently submitted to BIAS-2015 (v2.1.1), InterVar, and AIVA. A dedicated Assembly Stress-Test sub-study (n = 644 variants deliberately mislabelled genome builds) probed AIVA's input-error resilience. All outputs were evaluated against expert VCEP classifications using macro F1 scores.

Figure 3

Variant prioritization benchmark across 1,396 simulated rare disease cases. *(1,396 simulated rare disease patients; AIVA vs. LIRICAL vs. Exomiser). (A) Overall metrics table: Recall@K for K = 1, 5, 10, 20, and recall at Top-20 by category breakdown; highlighted cells indicate the best-performing tool per row. (B) Distractor-gene type distribution across the* 387 non-rank-1 cases; bars show the proportion of each distractor category (e.g., pathogenic-but-phenotype-irrelevant, phenotype distractor, Universal distractor, Tissue distractor, etc.). (C) Category-wise distractor breakdown (% of all patients per knowledge-graph coverage category = total error rate); stacked bars reflect the relative contribution of each distractor type per category.

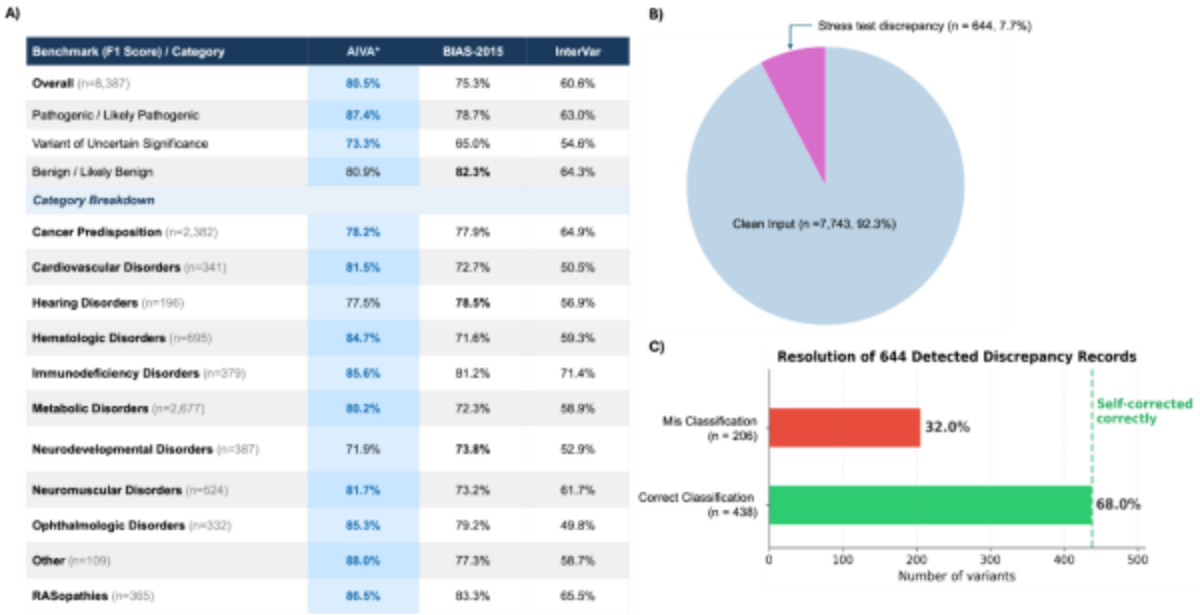

Figure 4

Variant classification performance and assembly self-correction stress test. (A) F1 score comparison of AIVA, BIAS-2015, and InterVar across overall performance, per-class breakdown (Pathogenic/LP, VUS, Benign/LB), and 11 disease categories; AIVA column highlighted in blue; bold values indicate the best-performing tool per row. (B) All 8,387 records stratified by assembly stress-test outcome: clean input (92.3%, n=7,743); self-corrected correctly (5.2%, n=438); self-corrected wrong (2.5%, n=206). (C) Resolution breakdown of the 644 deliberately mislabelled records: 438 self-corrected correctly (68%); 206 corrected wrong (32%). AIVA detected 100% of deliberately mislabelled inputs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfiles.docx