Folklore

PUBLIC SCIENTIFIC REPORT

Cohort 3: Real-World Performance Study

Expanded multi-source evaluation with independent per-variant forensic re-evaluation of classification differences

Document	HEL-RWP-C3-2026-001-PUBLIC
Version	1.0, August 2026
Study period	Cohort closed June 2026
Classifier	Folklore v3.39.1 at study closure
Author	Vladimir Mitev, Helena Bioinformatics Ltd.
ORCID	0009-0001-8013-6850
Report DOI	10.17605/OSF.IO/K8UWF
Public source	https://folklore.helena.bio/studies/cohort-3

This report contains aggregate, non-identifiable results only. It excludes patient registries, VCF files, case manifests, individual clinical reports, and per-case concordance records.

Suggested citation

Mitev V. Cohort 3: Real-World Performance Study. Helena Bioinformatics public scientific report, 1.0, August 2026. doi:10.17605/OSF.IO/K8UWF.

Abstract

Cohort 3 evaluated Folklore v3.39.1 in an expanded real-world programme. The manifest contained 60 cases; 57 had completed comparator data and were scored, three awaiting-comparator cases were excluded, and one pharmacogenetic re-analysis was segregated from ACMG per-variant scoring. The analytical denominator included 186 in-scope shared variants, of which 184 were detected, for 98.9 percent detection. The classification denominator included 184 shared outcomes: 102 FULL, 19 CLINICAL, 56 PARTIAL, and seven DISCORDANT, with 13 NOT EVALUABLE outcomes reported separately. FULL plus CLINICAL agreement was 65.8 percent. Every PARTIAL was an adjacent LP/VUS or LB/VUS difference, and all seven DISCORDANT outcomes were characterized as both-defensible rather than opposite-direction errors. Independent per-variant forensic re-evaluation found zero confirmed classifier-logic defects. AI-bearing reports were free of gene-disease name hallucination in 49 of 49 cases.

Objective

To characterize Folklore under a broader real-world case mix, quantify analytical and classification performance with explicit denominators, and independently re-evaluate every LP/VUS difference through the full classifier path.

Scope

This is a Phase 1 scientific performance report. It describes a bounded observational study and does not constitute regulatory clearance, a diagnostic release, a prospective multi-site trial, or authorization for unsupervised clinical use.

Terminology

Variant classes follow the five-tier ACMG/AMP framework: Pathogenic, Likely Pathogenic, Variant of Uncertain Significance, Likely Benign, and Benign [1]. FULL denotes identical class. CLINICAL denotes a one-tier difference within the Pathogenic and Likely Pathogenic boundary. PARTIAL denotes an adjacent boundary difference such as Likely Pathogenic versus VUS. DISCORDANT is reserved for more substantial disagreement. NOT EVALUABLE is reported separately from concordance denominators.

1. Background

Clinical variant classification combines population, computational, functional, segregation, allelic, and disease-specific evidence. The ACMG/AMP framework provides a common vocabulary and evidence structure, but qualified laboratories may still reach different classifications because of evidence availability, rule specification, transcript choice, disease mechanism, and the date of evaluation [1,3,4].

Published inter-laboratory studies show that structured review and data sharing can substantially improve concordance [3,4,6]. These findings support transparent characterization of differences rather than treating any single comparator as unquestioned ground truth. Folklore studies therefore distinguish analytical performance, classification performance, and clinical workflow performance, and they use multi-source evidence where the study design permits [7].

2. Methods

Study design

- Sixty manifest cases, with 57 scored after excluding three awaiting-comparator cases from every denominator.
- Fifty-six cases scoreable per variant after segregating one pharmacogenetic re-analysis from ACMG scoring.
- Coverage of more than 15 clinical domains, eight or more inheritance patterns, diagnostic indications, and genotype-only carrier-screen contexts.
- Analytical denominator of 186 in-scope shared variants and classification denominator of 184 shared outcomes.
- Independent forensic re-evaluation of LP/VUS differences against the live study databases and the complete classifier eligibility, guard, evidence, and combining path.

Comparator and ground truth

The peer reference laboratory served as a comparator. For real-world performance studies, comparator findings were interpreted alongside higher-weight public evidence, including ClinGen expert-panel curations and ClinVar records with review status preserved. The study did not assume that a single source was definitive in every disagreement.

Privacy-preserving analysis

All calculations reported here are cohort-level aggregates. No patient-level phenotype, genotype, sample identifier, referral date, clinical report, or rare case combination is reproduced. Small-cell narratives and individual variants are intentionally suppressed.

3. Results

Metric	Result	Interpretation
In-scope detection	184/186 (98.9%)	Wilson 95% CI 96.2-99.7
FULL concordance	102/184 (55.4%)	Identical five-tier class
FULL + CLINICAL	121/184 (65.8%)	Clinical-equivalent agreement
PARTIAL components	56	Across 34 cases
DISCORDANT outcomes	7	All both-defensible
NOT EVALUABLE	13	Reported outside fractions
Confirmed classifier defects	0	After forensic re-evaluation
AI faithfulness	49/49 (100%)	No gene-disease name hallucination

Selected performance metrics

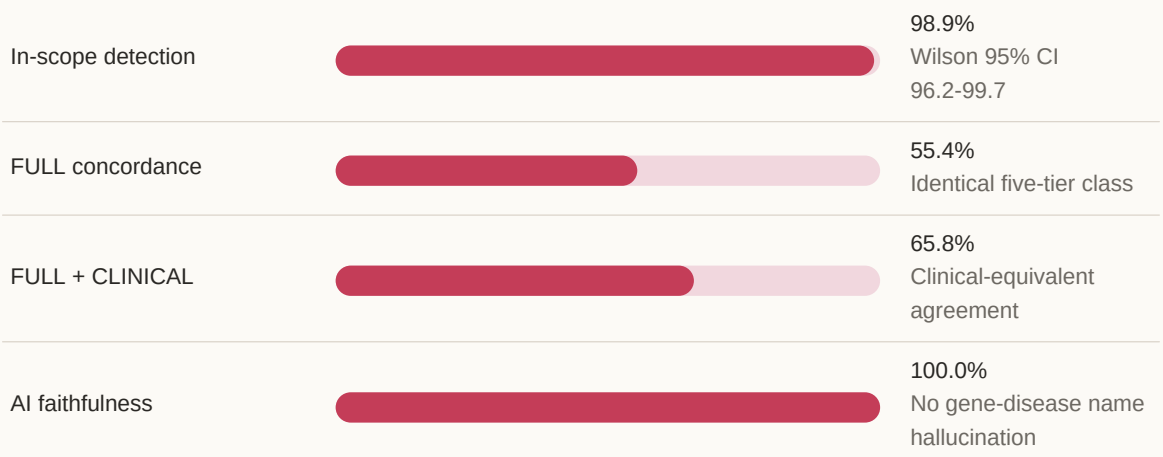

Figure 1. Selected aggregate percentages. Denominators are stated in Table 1.

4. Principal findings

Analytical performance

Folklore detected 184 of 186 in-scope shared variants. Two in-scope absences and 11 additional out-of-scope or annotation-boundary events were characterized separately rather than merged into the classification denominator.

Classification performance

FULL plus CLINICAL agreement was 121 of 184 shared outcomes. Every PARTIAL component occurred at an adjacent LP/VUS or LB/VUS boundary. No opposite-direction Pathogenic or Likely Pathogenic versus Benign or Likely Benign error was reported.

Forensic re-evaluation

The LP/VUS difference set was traced through evidence eligibility, conservative guards, evidence criteria, and combining rules. No classifier-logic defect was confirmed. The remaining differences reflected validated guards, correct eligibility outcomes, evidence gaps, or comparator over-calling of single carrier alleles.

Engineering findings

Two engineering improvements surfaced by the programme were deployed: re-enablement of PM5 evidence after validation of its data source, and widened compound-heterozygous splice-partner detection with the existing evidence guard retained.

AI workflow

All 49 AI-bearing reports were free of gene-disease name hallucination. Narrative faithfulness remained analytically separate from deterministic ACMG classification.

5. Discussion

Cohort 3 provides a larger and more heterogeneous evidence base than the preceding studies. Its value lies not only in the headline percentages but in the explicit denominator discipline and the independent re-evaluation of every important boundary difference.

The 55.4 percent FULL rate should not be interpreted without the 65.8 percent FULL plus CLINICAL rate and the disposition of all non-full outcomes. Published studies show that five-tier agreement is sensitive to rule specification, evidence age, and the LP/VUS boundary [3,4,6].

The absence of a confirmed classifier-logic defect after re-evaluation does not imply perfect platform performance. It means that the reviewed classification differences were explainable within the locked specification and available evidence. Analytical scope limits, report-generation issues, evidence gaps, and future validation needs remain relevant.

The separate AI faithfulness result is encouraging but bounded. It does not replace human review, does not alter the deterministic class, and does not establish performance outside the evaluated cases.

6. Limitations

- Observational case mix from a single peer comparator relationship.
- Three cases lacked completed comparator data and were excluded from denominators.
- Structural, copy-number, repeat-expansion, and some non-coding events were outside the evaluated classifier scope.
- The classification denominator contains shared comparator outcomes and does not score Folklore-only findings without a comparator.
- The study was not a prospective multi-site trial and used no independent external adjudication panel.
- AI faithfulness was assessed only in AI-bearing reports and should not be generalized beyond this cohort.

7. Data availability and privacy

The public data package contains this report, an OSF study summary, and links to the public Folklore evidence page. Patient registries, selection workbooks, VCF or alignment files, individual clinical reports, case manifests, and per-case concordance records are not publicly deposited. Genomic data is inherently identifying, and removal of names alone does not make a clinical genomic record anonymous.

Restricted material may be considered for controlled review only when a lawful basis, purpose limitation, confidentiality terms, an appropriate data-sharing agreement, and a secure transfer mechanism are in place. Access is assessed case by case and is not guaranteed.

8. Conflict of interest and provenance

- The author is founder, managing director, and shareholder of Helena Bioinformatics, which develops Folklore.
- The study used a peer laboratory with which Helena Bioinformatics has a professional collaboration. Multi-source evidence and independent per-variant forensic review were used to reduce single-source dependence.
- No external adjudication panel was engaged at this Phase 1 stage.

9. Ethics and responsible interpretation

This public report is a secondary aggregate account of work performed under collaboration and confidentiality arrangements. It includes no identifiable participant material and introduces no new participant intervention. The report must be read together with its stated limitations and must not be used as an individual diagnostic result.

10. References

1. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30.

2. Roy S, Coldren C, Karunamurthy A, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. J Mol Diagn. 2018;20(1):4-27. doi:10.1016/j.jmoldx.2017.11.003.

3. Amendola LM, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories. Am J Hum Genet. 2016;98(6):1067-1076. PMID:27181684.

4. Amendola LM, Muenzen K, Biesecker LG, et al. Variant classification concordance using the ACMG-AMP guidelines across nine genomic implementation research studies. Am J Hum Genet. 2020;107(5):932-941. doi:10.1016/j.ajhg.2020.09.011.

5. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(12):2163-2177. doi:10.1016/j.ajhg.2022.10.013.

6. Mighton C, Smith AC, Mayers J, et al. Data sharing to improve concordance in variant interpretation across laboratories. J Med Genet. 2022. PMID:33875564.

7. Mitev V. A methodological framework for real-world performance studies of clinical variant classification platforms at early organizational stages. Zenodo. 2026. doi:10.5281/zenodo.21105027.

Public research links

Study page: https://folklore.helena.bio/studies/cohort-3

OSF validation collection: https://osf.io/xnfzd/

Report DOI: https://doi.org/10.17605/OSF.IO/K8UWF

Parent project DOI: https://doi.org/10.17605/OSF.IO/8QBMH

Contact: https://www.helena.bio/contact