Folklore

PUBLIC SCIENTIFIC REPORT

Cohort 4: Deterministic Classification Performance Study

Large-scale fixed-version evaluation with session-level forensic review of material classification differences

Document	HEL-RWP-C4-2026-001-PUBLIC
Version	1.0, August 2026
Study period	Cohort closed July 2026; forensic review completed August 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/NPHW8
Public source	https://folklore.helena.bio/studies/cohort-4

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 4: Deterministic Classification Performance Study. Helena Bioinformatics public scientific report, 1.0, August 2026. doi:10.17605/OSF.IO/NPHW8.

Abstract

Cohort 4 evaluated Folklore v3.39.1 across 100 completed real-world cases, the largest fixed-version cohort in the programme. The study scored 395 comparator outcomes; 28 NOT EVALUABLE outcomes were reported separately and 367 shared outcomes formed the classification denominator. The denominator contained 223 FULL, 23 CLINICAL, 113 PARTIAL, and eight DISCORDANT outcomes. FULL agreement was 60.8 percent and FULL plus CLINICAL agreement was 67.0 percent, within the published inter-laboratory range used by the programme. All PARTIAL outcomes were adjacent LP/VUS or LB/VUS differences. The eight DISCORDANT outcomes concentrated in risk, hypomorphic, or haplotype-dependent alleles. Session-level forensic review of 19 priority variants across 16 cases confirmed four variant-level defects across three deterministic mechanisms and one output-architecture defect family. The study converted these findings into a defined correction and full-cohort regression programme.

Objective

To characterize deterministic ACMG classification performance at larger scale, distinguish classifier defects from reference-data gaps and legitimate methodological variation, and define a verifiable correction programme.

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

- One hundred real-world cases completed under one locked classifier version; no session failed.
- A classification denominator of 367 evaluable shared outcomes from 395 total comparator outcomes.
- Peer comparison interpreted alongside ClinGen expert specifications, ClinVar review status, population evidence, calibrated predictors, disease mechanism, and available functional or clinical evidence.
- Every PARTIAL, DISCORDANT, and NOT EVALUABLE outcome received a methodological disposition.
- Nineteen priority variants across 16 cases were reconstructed from stored criteria, frequencies, mechanism records, quality fields, guards, and combining rules.

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
Completed cases	100/100 (100%)	No failed session
Evaluable outcomes	367/395 (92.9%)	Classification denominator
FULL concordance	223/367 (60.8%)	Wilson 95% CI 55.7-65.6
FULL + CLINICAL	246/367 (67.0%)	Wilson 95% CI 62.1-71.6
PARTIAL outcomes	113/367 (30.8%)	All adjacent boundaries
DISCORDANT outcomes	8/367 (2.2%)	Risk or haplotype concentrated
NOT EVALUABLE	28	Reported outside fractions
Priority forensic set	19 variants	Across 16 cases
Confirmed defect findings	4 + 1 family	Variant-level plus architecture

Selected performance metrics

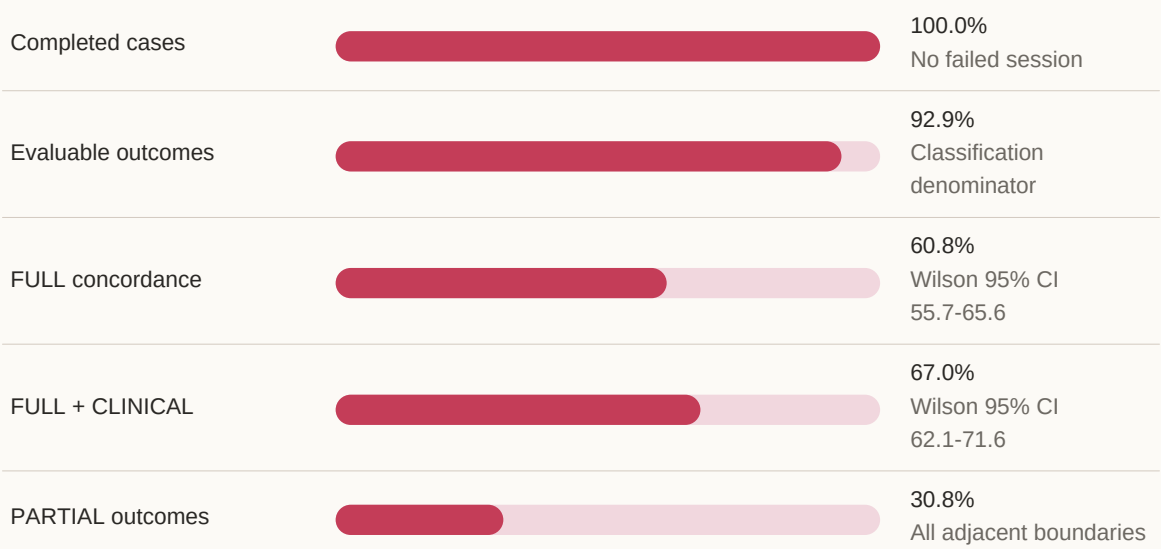

Figure 1. Selected aggregate percentages. Denominators are stated in Table 1.

4. Principal findings

Classification performance

FULL plus CLINICAL agreement was 246 of 367 evaluable outcomes. All 113 PARTIAL outcomes occurred at adjacent LP/VUS or LB/VUS boundaries. The eight wider differences concentrated in penetrance-, risk-, and haplotype-dependent alleles rather than a general ACMG combining-rule failure.

Deterministic defects

The forensic review confirmed disease- and inheritance-specific BS1 selection errors, incomplete dual-mechanism representation affecting PVS1 eligibility, and incorrect BP4_Moderate strength handling. A separate architecture family allowed case-level carrier status to overwrite the variant ACMG class.

Reference and manual evidence

Seven loss-of-function mechanism records were confirmed as missing or incomplete and entered curated-data remediation. Other peer classifications depended on segregation, phase, functional, phenotype, or literature evidence unavailable to automated VCF-only classification.

Correction programme

The confirmed defects will be corrected through disease-aware frequency selection, dual-mechanism modeling, calibrated evidence-strength mapping, and separation of variant class from carrier status. All 100 cases will then be reprocessed in shadow mode against the locked baseline.

5. Discussion

Cohort 4 was successful. It completed the programme's largest fixed-version study, placed aggregate peer agreement within the expected inter-laboratory range, and localized the material deterministic failures to a bounded set of mechanisms.

The result separates three kinds of work. Logic defects require code changes. Missing disease-mechanism records require versioned reference curation. Differences driven by segregation, phase, functional studies, or penetrance require qualified interpretation or a dedicated reporting model rather than invented automated evidence.

The correction gate preserves the current cohort as a baseline. Release requires the expected transitions on every confirmed defect, no regression among the 246 FULL or CLINICAL anchors, and no new unexplained opposite-direction result after full-cohort shadow reprocessing.

Clinical presentation findings are tracked separately from deterministic classification. Quality-failed calls must not become primary narrative findings, and genotype-only screens must remain distinct from phenotype-led diagnostic interpretation.

6. Limitations

- Single peer comparator relationship; multi-source review does not replace a multi-laboratory study.
- Descriptive observational case mix rather than a prospective randomized or pivotal cohort.
- Incomplete ClinGen expert-panel and gene-disease mechanism coverage across rare-disease genes.
- Some evidence types require manual segregation, phase, functional, phenotype, or literature review.
- Many screening cases lacked encoded phenotype terms, limiting phenotype-ranking assessment.
- No independent external adjudication panel was used at this Phase 1 stage.

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 session-level forensic review reduced single-source dependence.
- No external adjudication panel was engaged at this Phase 1 stage. The correction programme requires trigger-level and full-cohort regression evidence before release.

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-4

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

Report DOI: https://doi.org/10.17605/OSF.IO/NPHW8

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

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