Additional file 1

Methods, figures and tables

Intrinsic sequence evidence cannot reach a pathogenic classification for missense variants under a partitioned ACMG framework

Contents

Methods

Figure S1. A worked example: the safety interlock on a single case

Figure S2. Sensitivity of the diagnosis arm to the strength of the gene term

Figure S3. Treatment-divergence safety interlock, by scenario

Figure S4. Diagnosis-posterior calibration on the curated benchmark

Figure S5. Independent sensitivity on the CDC Hemophilia Mutation Projects

Table S5. Per-criterion agreement with expert-panel applications

Table S6. Attribution of the intrinsic-evidence ceiling

Table S7. Treatment-divergence scenarios

Table S8. LIRICAL comparison arms

Table S9. CHAMP and CHBMP recall by gene

Tables S1, S2, S3 and S4 are supplied as separate Excel files, being large data tables. Every figure and table here is also archived, with the scripts that generate them, in the data record.

Methods

Variant Curation Expert Panel specifications as implemented

Gene-specific rule sets are encoded as versioned YAML, one file per specification, loaded at run time rather than compiled into the scoring logic. Four elements are captured per specification: the frequency thresholds that determine BA1, BS1 and PM2; any modification to a criterion's default strength; the computational thresholds that determine PP3 and BP4; and the identifier of the specification itself, so a classification can be traced to the rule set that produced it.

Frequency criteria. Thresholds were taken from the ClinGen Criteria Specification Registry where a specification exists, and cross-checked against the allele frequencies curators cite in the corresponding Evidence Repository records. This cross-check corrected several values during development and gave 97.8% concordance on the 629 variants for which a curator-cited frequency was available.

Strength modification. Where a panel modifies a criterion's default strength, the modification is encoded explicitly. The most consequential is PM2 at Supporting rather than the ACMG default of Moderate, which several panels specify and which materially changes the point total for rare missense variants, PM2 being the most frequently applied criterion in this domain.

Computational thresholds. Each specification carries its own REVEL and SpliceAI cut-offs for PP3 and BP4 where the panel defines them. Where a panel does not, the ClinGen-calibrated bands are used as a fallback. The RUNX1 specification uses the Myeloid Malignancy panel's version 2 values (PP3 at REVEL >= 0.88, BP4 at REVEL <= 0.50) rather than generic defaults.

Residual simplifications, stated. Two elements of the published specifications are not fully encoded. PVS1 is implemented as the Abou Tayoun decision tree rather than as each panel's gene-specific elaboration of it, and PS4 is implemented as a generic odds-ratio and proband-count tree rather than each panel's semi-quantitative counting rules. Both are documented simplifications rather than omissions; neither affects the results reported here, since PS4 is never applied for want of case-control input and PVS1 does not apply to the missense surface that carries the analysis.

Implemented in rules/vcep/loader.py and rules/vcep/specs/.yaml.*

The evidence partition

Each ACMG/AMP criterion is mapped to exactly one owning factor by a static table. Criterion strings are normalised before lookup, so a strength-modified form such as PM2_Supporting resolves to the same owner as PM2. The mapping is total over the criterion vocabulary encountered in the Evidence Repository: the genome-wide analysis reports 100.0000% coverage over 38,802 applied criteria with no unrecognised code.

Exclusivity is enforced by construction rather than by convention, and verified by a test that supplies the joint model a variant with the criteria PP4, PS3, PP1 and PM3 added explicitly and asserts the variant marginal is unchanged to within 1e-12. Because those criteria are owned by other factors, re-adding them to the genetic stream must have no effect; the test fails if the partition is ever weakened.

Implemented in rules/vcep/partition.py; verified by tests/test_discern_joint.py.

The gene term and its two bounds

The joint factorises as

$$P(D, V \mid E) \propto P(E_{pheno} \mid D) \cdot P(E_{geno} \mid V) \cdot P(E_{func} \mid D, V) \cdot P(V \mid G, D) \cdot P(G \mid D) \cdot P(D)$$

P(V | G, D) normalises over the five variant states, so it cancels when the joint is marginalised to a disease posterior. Without the separate P(G | D) term the gene therefore carries no information about disease identity, which is the defect described in the main text. P(G | D) is implemented as an on-gene/off-gene likelihood pair, giving a likelihood ratio of 4.

Two bounds constrain that value, and both are architectural rather than fitted:

1. **It must not exceed a cluster's deciding observation.** The sharpest discriminator in the knowledge base is the ristocetin-induced platelet aggregation mixing study, at a likelihood ratio near 11. If the gene term exceeded that, no laboratory result could overturn the gene, and the abstention rule could never name a next test that changes the leading diagnosis. A ratio of 4 is the largest tested value at which the mixing study still overturns the gene.

2. **It must not veto a treatment-divergent competitor.** Finding a variant in one gene does not exclude a disease of another gene, which may never have been sequenced or whose variant may be benign. The safety interlock is therefore evaluated on a gene-blind posterior, obtained by omitting P(G | D), while the leading diagnosis reported to the user remains the gene-aware one.

A sweep across the full range of the parameter (Figure S2) shows hard-stop sensitivity and specificity invariant at 100% throughout, the curated diagnosis result flat above the committed value, and the deciding-assay bound failing at ratios of 9 and above.

Implemented in jointdx/factorgraph.py; swept by bench/phase_r_gene_term_sensitivity.py.

Calibration and its out-of-sample estimation

Variant scores are calibrated by isotonic regression. Estimation is out-of-fold throughout: a 5-fold stratified split with shuffling and a fixed seed is constructed once per surface; the isotonic map is fit on the four training folds and applied to the held-out fold; expected calibration error and Brier score are accumulated on held-out predictions only. No variant is ever scored by a calibrator that has seen its label.

Three properties are asserted in code at run time rather than checked afterwards: that no fold trains and evaluates on the same index, that no index appears in two evaluation folds, and that every index is held out exactly once. Any violation raises rather than returning a number.

The fold assignment itself is published (benchmarks/variant_arm/calibration_folds.json in the data record), listing for every variant which fold held it out, together with the full training and evaluation index sets. A reader can verify the three properties directly rather than taking the protocol on trust.

Expected calibration error uses ten equal-width bins over the unit interval, weighting each bin's absolute gap between mean predicted probability and observed frequency by that bin's occupancy. The final bin includes probability 1.0.

Comparator calibration. REVEL and AlphaMissense were put through the identical fold assignment and the identical isotonic procedure, and are reported both raw and calibrated. This is what makes the calibration comparison like-for-like, and it is why no claim of unique calibration is made.

Implemented in bench/phase_r_variant.py; the calibration primitive is core/stats.ece.

Statistical procedures

Confidence intervals on proportions. Wilson score intervals. Preferred to the normal approximation because the reported rates sit far from one half at large sample size, where Wald intervals are unreliable.

Confidence intervals on metrics. Percentile bootstrap, 1,000 resamples, fixed seed. Resamples in which a class is absent are discarded rather than scored. Intervals are the 2.5th and 97.5th percentiles of the resampling distribution.

Comparing two areas under the curve. DeLong's test in the fast form of Sun and Xu, computed on the variants both tools score, with ties handled by mid-ranks. Reported alongside a paired bootstrap on the difference, which resamples variants and rescores both tools on the same resample, so the interval reflects the pairing.

Comparing two classifiers on the same cases. Exact McNemar, using the binomial test on the discordant pairs. Chosen over the chi-squared approximation because the discordant counts here are small.

Rank metrics. Recall at k and mean reciprocal rank from one-based hit positions, with a non-hit contributing zero to the reciprocal rank.

Every one of these is implemented once, in a module with no dependency beyond numpy and scipy, and is covered by tests against closed forms and hand-computed values. They were extracted from the analysis harnesses precisely so that the arithmetic behind the reported intervals could be exercised by continuous integration rather than assumed.

Implemented in core/stats.py; verified by tests/test_stats.py.

The ClinVar-blinded protocol, per tool

Several comparators consume ClinVar through the PP5 and BP6 criteria, which would let them reproduce rather than predict the label on any ClinVar-derived truth surface. The protocol therefore differs by tool and is stated for each.

DISCERN applies no ClinVar-derived criterion. PS1 and PM5, which require a same-residue ClinVar lookup, are implemented but deliberately not wired in for these runs. This is the reason they appear zero times in the per-criterion comparison, and it is a protocol choice rather than a gap in the engine.

REVEL and AlphaMissense are precomputed scores and consume no classification.

GeneBe does apply PP5 and BP6. Rather than exclude it, it is retained as an exhibit: it attains a perfect area under the curve on the eRepo surface and continues to do so under the time-split, which demonstrates the circularity concretely.

InterVar and BIAS-2015 are positioned by published metrics obtained on different variant sets, and are labelled as such wherever they appear.

The time-split panel additionally restricts grading to variants whose expert classification post-dates the ClinVar snapshot bundled with the comparator tools (2021-05-01), so that no tool could have memorised the label irrespective of protocol.

The curated case benchmark and the HPO crosswalk

Cases were curated from published reports, each represented as the causal gene together with the clinical and laboratory findings the report describes, and each carrying the PubMed identifier of its source. No article text is reproduced anywhere in the deposit; the benchmark ships as identifiers, extracted terms, and the expected diagnosis.

Representability in the Human Phenotype Ontology was determined by inverting the committed crosswalk, which maps HPO terms to the engine's internal finding vocabulary, and asking which findings used anywhere in the benchmark have at least one HPO term. Thirteen of the forty-eight distinct findings do. The ontology carries generic terms for several of the assay families involved, impaired ristocetin-induced aggregation among them, but the thirty-five findings that fail the test are discriminating results it does not represent at that resolution, among them the ristocetin mixing study that separates plasma from

platelet origin, the urokinase overexpression assay, and the prothrombinase assay. Nineteen of the forty-two cases carry no HPO term at all and cannot be ranked by a phenotype-driven tool.

Implemented in eval/phenotype_tool_comparison.py; verified by tests/test_diagnosis_baselines.py.

The phenotype-blind baselines

Three reference points, none of which sees a phenotype:

Gene lookup. Ranks the cluster by placing the diseases that list the case's gene first, each block ordered by prior. A static table with no likelihood ratios and no findings. Verified to produce an identical ranking when every finding is stripped from the case.

Prior only. Ranks the cluster by prior and ignores the gene entirely.

Uniform random within cluster. The floor, reported as the analytic expectation together with a simulated interval.

Cases are stratified into three mutually exclusive groups: those whose gene maps to exactly one disease within its cluster, those carrying no causal gene, and the informative subset whose gene is present but maps to more than one disease, so that something other than the gene must break the tie. The three groups partition the benchmark exactly.

Implemented in eval/gene_only_baseline.py.

LIRICAL run configuration

LIRICAL version 2.4.1 was run in phenotype-only mode inside an eclipse-temurin:17-jre container. Nothing was installed on the host.

Per case, the observed HPO terms were supplied via --observed-phenotypes and, where the case records an explicitly absent finding with an HPO representation, the corresponding terms were supplied via --negated-phenotypes, so that LIRICAL receives the same pertinent negatives the engine uses. Genome build was left unset, which selects phenotype-only operation and requires no VCF. Only the twenty-three cases carrying at least one HPO term were run; the remaining nineteen cannot be ranked by a phenotype-driven tool and are reported as such rather than scored as failures.

A practical note for reproduction. LIRICAL's own downloader retrieves mim2gene_medgen over FTP, which many institutional networks block. The same file is served over HTTPS from the identical NCBI path, and substituting it allows the remaining resources to download normally.

Scoring. LIRICAL ranks diseases genome-wide by OMIM and Orphanet identifier, whereas the engine ranks within a cluster of three to eight. Scoring is therefore performed at gene and disease-family resolution: a hit counts if the top-k contains any disease identifier the Human Phenotype Ontology annotates to the case's causal gene. The gene-to-disease crosswalk is derived from the committed HPO annotation file rather than asserted, so every identifier traces to a public source. This is the generous reading for LIRICAL, since it cannot be penalised for differences in OMIM granularity.

A second, like-for-like arm collapses LIRICAL's genome-wide ranking onto the same cluster members the engine ranks, so both tools order the same three to eight diseases. That arm is the one the main text quotes.

Exomiser was not run. It requires a VCF, and seeding one per case with zygosity matched to each report's stated inheritance introduces a correctness risk judged larger than the comparison's marginal value once LIRICAL had established the phenotype-channel result on the same inputs. This is recorded as a scope decision rather than presented as a handicapped run.

Implemented in eval/lirical_arm.py.

Reproducibility

All reported values are produced by scripts archived with the code record and are locked by regression tests that read the committed outputs, so a change to any benchmark that moves a reported number fails the build rather than propagating silently. A further test scans the documentation for a list of claims this analysis does not support and fails the build if any of them is asserted.

Third-party databases are referenced by versioned manifest giving source URL, version or snapshot date, and checksums of the exact files used, rather than redistributed. Software and database versions are listed in Table S4.

Figure and table legends

Figure S1. A worked example: the safety interlock on a single case. (A) The evidence supplied: a GP1BA missense variant absent from gnomAD, enhanced low-dose ristocetin-induced platelet aggregation, a mixing study of platelet origin, plasma origin explicitly excluded, thrombocytopenia, and desmopressin as the planned therapy. (B) The criterion trail. PM2 is applied and owned by the variant factor. PP4 and PS3 are owned by the disease and functional factors and are not available from sequence; PM5 and PS4 are variant-intrinsic but have no input under this protocol. The variant therefore remains of uncertain significance. (C) The ranked differential, in which platelet-type von Willebrand disease leads with a posterior mean of 99.6%; bars carry the 95% credible interval from Monte-Carlo resampling of the source frequencies. Panel D reports 99.7% for the same disease because the decision layer uses the analytic estimate at the nominal frequencies rather than the mean over resampled ones. (D) The decision output. Desmopressin is contraindicated in type 2B von Willebrand disease, a VWF disease, whereas the variant here is in GP1BA. Because safety is adjudicated gene-blind, type 2B retains 0.9% in the safety view and the hard stop fires, together with the sequencing test that would resolve the question. This is the framework's clinical argument on one case: a low-probability but treatment-divergent competitor is not silently retired by the gene. Related to Figure 1.

Figure S2. Sensitivity of the diagnosis arm to the strength of the gene term. Curated Top-1 accuracy, hard-stop sensitivity and hard-stop specificity across the full range of the P(G|D) likelihood ratio, from an inert gene term (ratio 1) to a ratio of 19. Safety behaviour is invariant throughout. The diagnosis result is flat above the committed value of 4, which is the largest ratio at which a cluster's deciding observation can still overturn the gene; shaded regions mark ratios at which it cannot. The vertical axis begins at 70% so that the range the three series occupy is legible. Related to Figure 6 and the Methods.

Figure S3. Treatment-divergence safety interlock, by scenario. Each of five treatment-divergent scenarios against the two behaviours the interlock must exhibit: firing when the planned management is contraindicated for a disease that has not been excluded, and remaining silent when the planned management is harmless. Sensitivity and specificity are both 100%. Adjudication is performed on a gene-blind posterior against the engine's own leading call, so that a variant in one gene cannot silently retire a treatment-divergent disease of another. Related to the trustworthiness results.

Figure S4. Diagnosis-posterior calibration on the curated benchmark. Mean predicted confidence against observed accuracy on the curated case set (n = 42), with the expected calibration error. The shaded band marks the region above the 0.8 confidence threshold, in which no incorrect call was made. Reported as a proof-of-concept characterisation: with the diagnosis arm at ceiling on this benchmark the calibration figure reflects mild underconfidence rather than a discriminating result, and the abstention layer's operational value requires a benchmark that is not at ceiling. Related to the abstention result.

Figure S5. Independent sensitivity on the CDC Hemophilia Mutation Projects. Likely-pathogenic or pathogenic recall on null variants in the CHAMP and CHBMP datasets (2,130 null variants of 5,437 disease alleles), scored by consequence alone with no frequency or predictor input. The lower F9 figure reflects that gene's large terminal exon, where premature termination codons escape nonsense-mediated decay and are correctly held at uncertain significance rather than called pathogenic. Related to the independent sensitivity result.

Table S1. Discrimination likelihood ratios for the ten confusable-disease clusters. Every likelihood ratio in the disease-discrimination model, given as the probability of the finding under each disease, with the sample size and the PubMed identifier of the primary source. Continuous integration fails if any entry lacks a source. Supplied as a separate Excel file.

Table S2. The curated published-case benchmark. All 42 cases with PubMed identifier, causal gene, cluster, expected diagnosis, extracted Human Phenotype Ontology terms, stratum, and the leading call and

Top-1 outcome for DISCERN and for the phenotype-blind gene lookup. No article text is reproduced. Supplied as a separate Excel file.

Table S3. Third-party data sources. Each source with the files used, the access URL, the version or snapshot date, and its role. None is redistributed; checksums for the exact files used appear in the deposit manifest. Supplied as a separate Excel file.

Table S4. Software, tool, and database versions. Every component used to produce the reported results, with its version and role. DISCERN's own commit hash is the authoritative version identifier. Supplied as a separate Excel file.

Table S5. Per-criterion agreement with expert-panel applications. Every ACMG/AMP criterion encountered on the Evidence Repository surface, with the number of times the expert panels applied it, the number of times DISCERN applied it, Cohen's kappa where both applied it at least once, the number of variants on which both applied it, and the fraction of those on which both applied it at the same ClinGen strength. Computed on the 1,265 variants carrying a pathogenic or benign expert classification, since agreement requires a graded variant on both sides; each kappa reconciles with the applied counts in its own row at that denominator. An unmodified code is resolved to its criterion's default strength before the strength comparison. Criteria applied zero times carry the reason, in the same three categories Table S6 uses: a protocol choice, a data-availability limit, or an engine scope gap.

Table S6. Attribution of the intrinsic-evidence ceiling. Variants reaching Likely Pathogenic under each restoration condition: as scored; restoring the variant-intrinsic criteria this pipeline cannot annotate; restoring the criteria the partition routes to other factors; and restoring both. Followed by the reason each unapplied criterion has no input, classified as a protocol choice, a data-availability limit, or an engine scope gap. Counts use default criterion strengths and are therefore lower bounds.

Table S7. Treatment-divergence scenarios. Each safety scenario with the disease pair, the planned management, the expected behaviour, and the observed behaviour, for both the contraindicated and the harmless plan.

Table S8. LIRICAL comparison arms. Every arm of the external comparison with its sample size, recall at 1, 3 and 5, and mean reciprocal rank, followed by the paired tests against LIRICAL restricted to the same cluster. The hpo_representable_only arm supports claims about inference on identical evidence; the phenotype_only arm supports the architectural claim about what the model can represent. The post-correction arm is in-sample and is not a head-to-head result.

Table S9. CHAMP and CHBMP recall by gene. Likely-pathogenic or pathogenic recall on null variants for F8, F9, and the combined datasets, with the explanation of the F9 result.

Figures

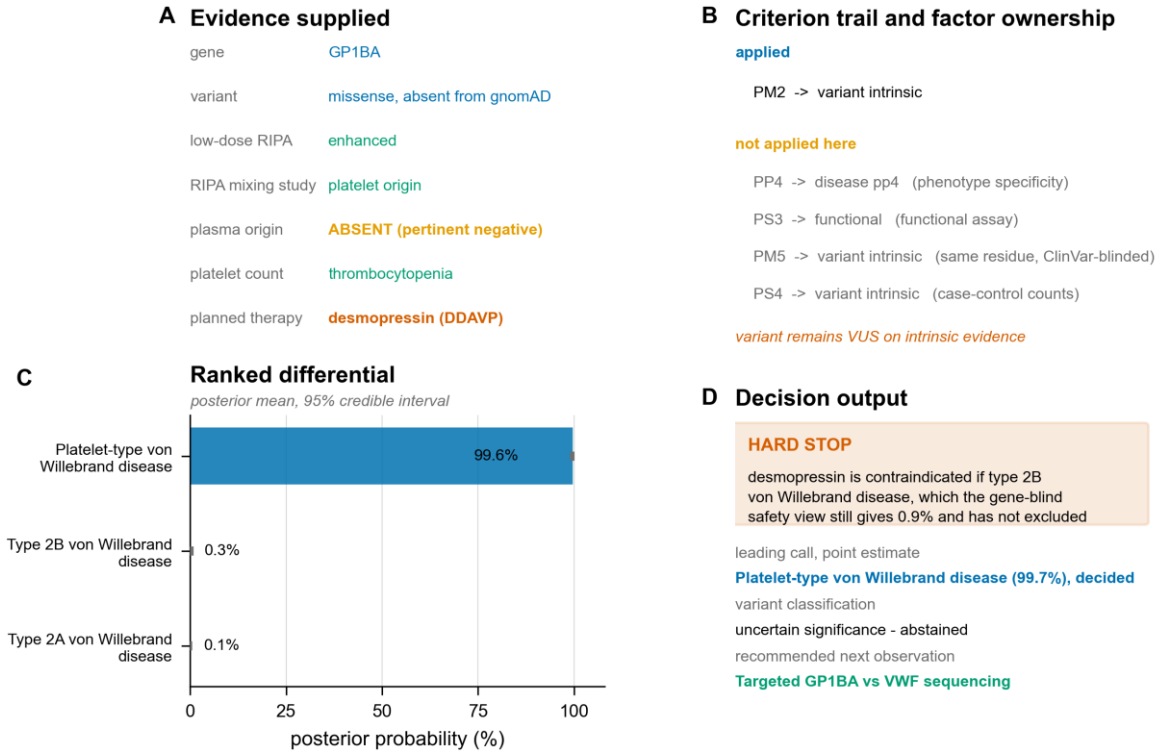

Figure S1. A worked example: the safety interlock on a single case.

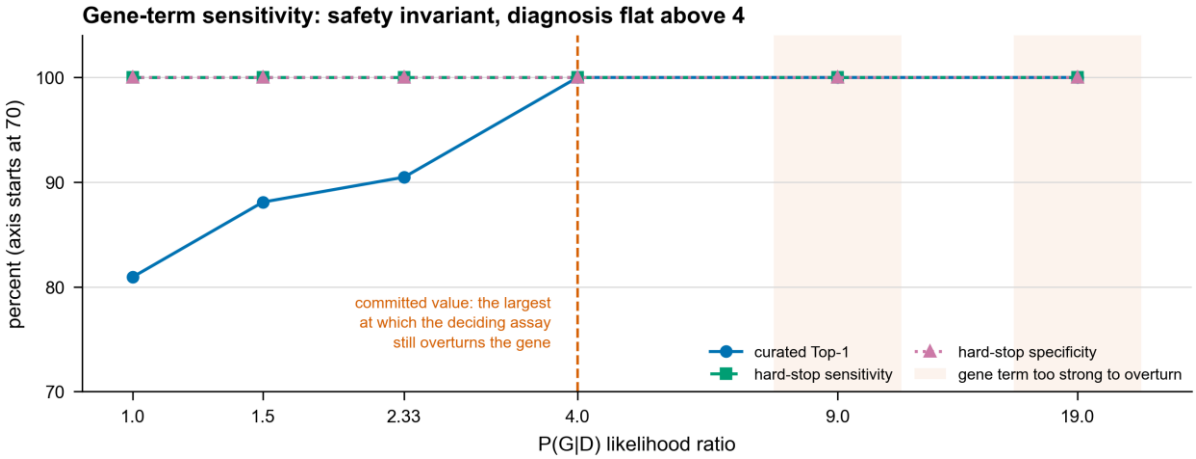

Figure S2. Sensitivity of the diagnosis arm to the strength of the gene term.

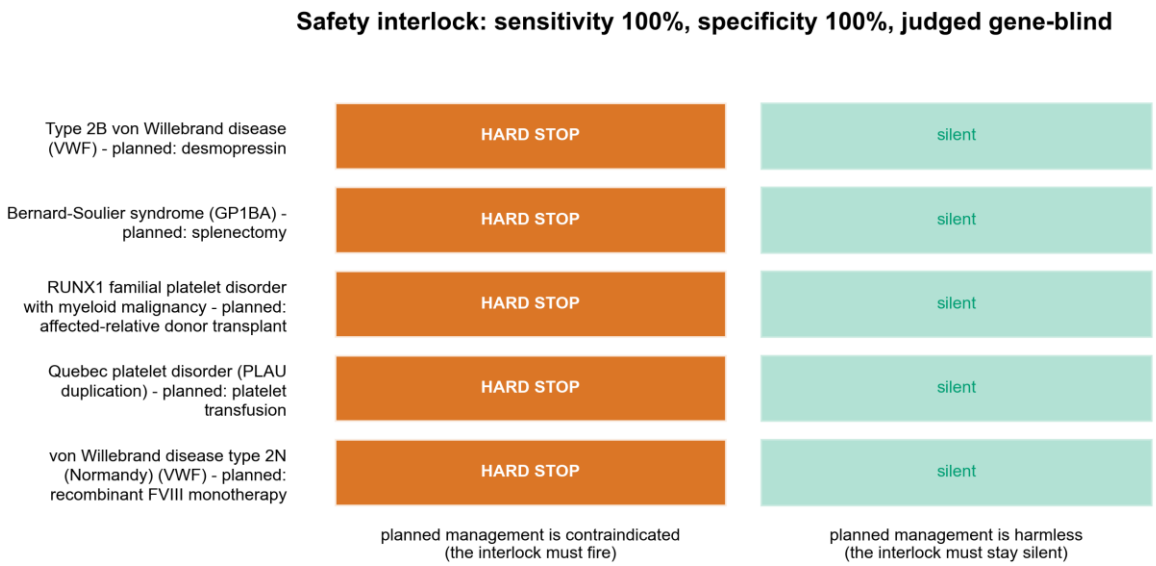

Figure S3. Treatment-divergence safety interlock, by scenario.

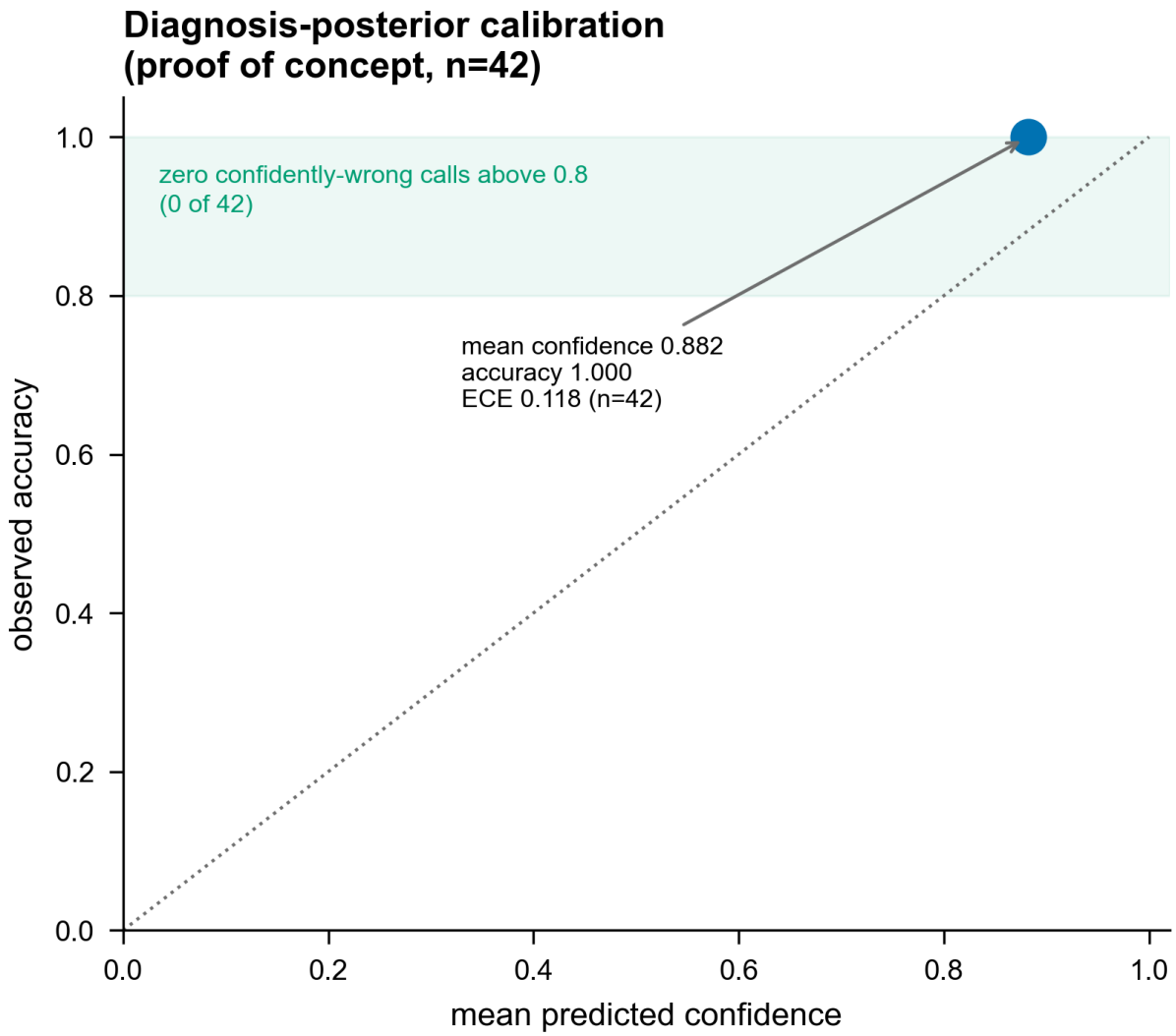

Figure S4. Diagnosis-posterior calibration on the curated benchmark.

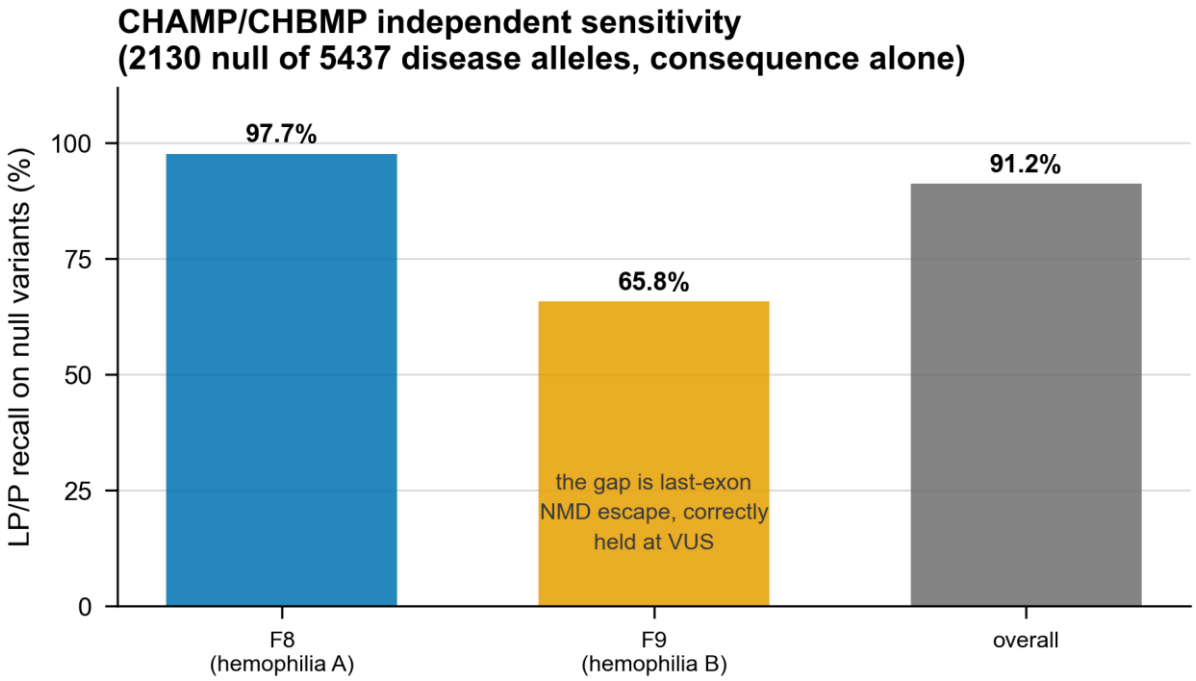

Figure S5. Independent sensitivity on the CDC Hemophilia Mutation Projects.

Tables

Table S5. Per-criterion agreement with expert-panel applications

ACMG criterion	erepo applied	discern applied	kappa	both applied	strength agreement	zero application cause	note
PVS1	178	173	0.977	172	0.041	applied by both	the criterion agrees almost perfectly; the strength does not, because the PVS1 decision tree needs transcript geometry (NMD prediction and the fraction of protein removed) that this annotation surface does not carry, so it downgrades to Moderate. Where those inputs are available, as in the CHAMP and CHBMP analysis, PVS1 reaches Very Strong
PP3	305	293	0.947	287	0.969	applied by both	derived from the same predictor scores the panels cite, at the ClinGen thresholds
BA1	183	98	0.64	95	1.0	applied by both	frequency-based; DISCERN is the more conservative of the two
BS1	85	58	0.297	24	1.0	applied by both	frequency-based; DISCERN is the more conservative of the two
PM2	812	479	0.084	336	0.94	applied by both	the panels apply it more often than DISCERN does (812 against 479); DISCERN requires the gnomAD frequency to be below the gene-specific threshold, and the annotation surface leaves that undetermined for many variants
BP4	615	91	0.074	67	1.0	applied by both	the panels apply it more often than DISCERN does (615 against 91); DISCERN applies it only at the ClinGen benign-side threshold
BP7	541	0		0		engine scope gap: applies to synonymous variants, outside the missense scope here	
BS2	5	0		0		engine scope gap: observation in healthy adults not implemented	
PM1	53	0		0		engine scope gap: hotspot and critical-domain annotation not implemented	
PM5	97	0		0		protocol choice: same-residue ClinVar lookup disabled by the blinding protocol	
PS1	7	0		0		protocol choice: same-residue ClinVar lookup disabled by the blinding protocol	
PS3	131	0		0		protocol choice: routed to the functional factor by the partition	
PS4	173	0		0		data-availability limit: no public surface supplies case-control counts at this scale	

Computed on the 1,265 variants of the bleeding-panel surface carrying a pathogenic or benign expert classification, not on all 2,239: agreement requires a graded variant on both sides. Every kappa here reconciles with the applied counts in the same row at that denominator. Criteria applied by neither the expert panels nor DISCERN are omitted (PS2, PM3, PM4, PM6, PP1, PP2, PP4, PP5, BP1, BP2, BP3, BP5, BP6), so these thirteen rows are not the whole of ACMG/AMP. Agreement is reported at two levels: kappa asks whether the criterion was applied at all, and strength_agreement asks, among variants where both applied it, how often they applied it at the same ClinGen strength. An unmodified code is resolved to its criterion's default strength before comparison.

Table S6. Attribution of the intrinsic-evidence ceiling

restoration condition	variants reaching LP	of n
as scored (intrinsic evidence only)	0	425
+ intrinsic criteria with no input here	27	425
+ criteria the partition routes away	52	425
both restored	89	425
code	reason unapplied	kind of limit
PS1	implemented (adapters/clinvar.py) but deliberately not wired in: this benchmark runs ClinVar-blinded, and PS1 needs a same-amino-acid-change ClinVar lookup. Supplying it would reintroduce exactly the circularity the GeneBe exhibit demonstrates, so its absence here is a protocol choice, not a gap.	protocol choice (ClinVar-blinded)
PM5	same as PS1 - implemented, and withheld for the same ClinVar-blinding reason (different missense at the same residue).	protocol choice (ClinVar-blinded)
PS4	implemented as a decision tree in rules/variant_scoring.py, but it requires case-control inputs - proband counts against expectation, or an odds ratio with its confidence bound - and the eRepo annotation cache carries none of them. A data-availability limit.	data availability
PM1	genuinely not implemented: no scorer emits PM1, and no VCEP specification in rules/vcep/specs/ encodes hotspot or critical-domain regions for the in-scope genes. The in_functional_domain annotation that exists feeds the PVS1 tree, not PM1. This is a scope limitation and the one item on this list that is a genuine engine gap rather than a protocol or data constraint.	engine scope gap

Points for the restored codes use default ACMG strengths, because stripping a strength suffix is how the code vocabulary is compared. VCEPs frequently apply gene-specific strengths above the default, so every count here is a lower bound on what the expert evidence would actually contribute.

Table S7. Treatment-divergence scenarios

scenario	planned management	expected on contraindicated	observed on contraindicated	expected on harmless	observed on harmless
rv_vwd2b_ddavp	ddavp	fire	fired	stay silent	silent
rv_bss_as_itp	splenectomy	fire	fired	stay silent	silent
rv_runx1_as_itp	affected_related_donor_transplant	fire	fired	stay silent	silent
rv_quebec_platelets	platelet_transfusion	fire	fired	stay silent	silent
rv_vwd2n_as_hema	recombinant_fviii_monotherapy	fire	fired	stay silent	silent

sensitivity 100%, specificity 100% over 5 scenarios; adjudicated gene-blind against the engine's own leading call.

Table S8. LIRICAL comparison arms

arm	n	recall@1	recall@3	recall@5	MRR
LIRICAL_genome_wide	23	0.1304	0.2609	0.3478	0.215
LIRICAL_restricted_to_cluster	23	0.5652	0.9565	0.9565	0.7536

DISCERN_hpo_representable_only	23	0.7391	1.0	1.0	0.8623
DISCERN_phenotype_only_no_gene	23	0.913	1.0	1.0	0.9565
DISCERN_pre_gene_term_fix	23	0.8261	1.0	1.0	0.913
DISCERN_post_fix_IN_SAMPLE	23	1.0	1.0	1.0	1.0
paired test	arm	delta	CI95_low	CI95_high	McNemar_p
vs LIRICAL restricted	phenotype_only	0.3478	0.1304	0.5652	0.021484
vs LIRICAL restricted	hpo_representable_only	0.1739	-0.0446	0.4348	0.289062
phenotype-only arm is invariant to the gene-term fix	True				

LIRICAL 2.4.1, phenotype-only mode, observed and negated HPO terms supplied, run in a container. Quote the hpo_representable_only arm for inference claims and the phenotype_only arm for the architectural claim.

Table S9. CHAMP and CHBMP recall by gene

gene	LP P recall on null	note
F8	0.977	NMD-triggering nulls resolved by PVS1
F9	0.658	gap is the large terminal exon: last-exon PTCs escape NMD and are correctly held at VUS
overall	0.912	2130 null variants of 5437 disease alleles

Recall by consequence alone, with no frequency or predictor input. Source: docs/DISCERN_CHAMP_CHBMP_Benchmark_v1.md