Embedding-Based Semantic Alignment Between 19th-Century Homeopathic Provings and Modern Clinical Toxicology: A Comprehensive Computational Analysis

Running title: Semantic Alignment of Historical Provings with Modern Toxicology

Author Information

Author: Gery Pollet[1]*

Affiliations: [1] Aquanta Labs — Independent Researcher

***Corresponding author:** gpollet@aquanta.com

Abstract

Background

In the 19th century, practitioners systematically recorded symptoms induced by pharmacological substances in healthy volunteers—a process known as “provings” in the homeopathic tradition. These historical observations constitute a large corpus of pre-modern clinical documentation. To our knowledge, limited prior work has quantitatively assessed whether these records align with modern medical knowledge at scale.

Objective

To determine whether historical proving descriptions of toxic substances show statistically significant semantic alignment with contemporary clinical toxicology signs.

Methods

We analyzed 820 remedies derived from pharmacologically active or toxic substances, comparing 126,667 symptom descriptions (74,415 unique texts) from six classical Materia Medica sources against a curated database of 4,091 modern clinical signs. Semantic similarity was computed using neural text embeddings (Qwen3 0.6B, 1024 dimensions). Our primary endpoint was top-3 unique-

ness-weighted similarity; three secondary statistics were also evaluated. Permutation testing (n=2,000) was performed under two null models, with Benjamini-Hochberg FDR correction applied globally and on a pre-specified high-evidence subset.

Results

Among 78 remedies (9.5%) meeting pre-specified quality criteria, 22 (28.2%) achieved significance after Bonferroni correction for multiple endpoints ($\alpha/4 = 0.0125$); 3 additional remedies were significant at per-endpoint FDR ($\alpha = 0.05$). Enrichment over chance was assessed exploratorily (Bonferroni tier: 22.6-fold, binomial $p = 9.8\times10^{-24}$). Negative control substances (lactose, ethanol, mineral water) showed no signal. Aligned remedies include Belladonna matching "dysphagia" (similarity 0.918), Antimonium tartaricum matching "thick white tongue coating" (0.901), and Plumbum metallicum matching "Burton's line" (0.796)—signs consistent with known toxidromes.

Conclusions

This comprehensive computational analysis reveals statistically significant semantic alignment between a subset of 19th-century homeopathic provings and modern clinical toxicology. These findings suggest that some historical descriptions may reflect recognizable toxicological patterns.

Keywords: homeopathy, toxicology, natural language processing, text embeddings, semantic similarity, historical medicine, validation study

Introduction

Homeopathy, developed by Samuel Hahnemann in the late 18th century, relies on "provings" (Prüfungen)—systematic experiments in which healthy volunteers ingest substances and record resulting symptoms [1]. These observations, compiled into reference texts called Materia Medica, form the empirical foundation of homeopathic prescribing. Provings were conducted over approximately two centuries by multiple physicians across diverse populations, with independent observers often recording concordant symptoms for the same substances [6,7,8,9]. While homeopathy remains controversial, with debates focusing on the plausibility of ultra-high dilutions [2], the historical proving observations themselves—as empirical recordings of physiological responses to pharmacologically active substances—have received limited scientific scrutiny.

Proving substances spanned a broad pharmacological range: industrial toxicants (lead, mercury, arsenic), plant alkaloids (belladonna, strychnine, conium), and biological substances (snake venoms, spider venoms, cuttlefish ink). Many of these have well-characterized toxidromes in modern clinical toxicology. If provers recorded their experiences with reasonable fidelity, these descriptions might show semantic overlap with modern toxicological knowledge. Conversely, if provings were purely subjective or fabricated, no systematic alignment would be expected.

An important methodological clarification: provers did not ingest raw toxic substances. Hahnemann prescribed the 30th centesimal potency (30C) as the standard proving dose (Organon §128 [1])—a dilution well beyond the Avogadro limit where no molecules of the original substance remain. The earliest provings (pre-1800) sometimes employed material doses such as tinctures or crude preparations, but the systematic provings compiled in Kent's Repertory and Hering's Guiding Symptoms drew on a mixture of material and highly diluted doses [6,8,9]. This creates an interpretive tension central to the present analysis: if provings conducted at ultra-high dilutions nonetheless produced symptom descriptions that align with modern toxicology, this alignment requires explanation—whether through pharmacological trace effects, observer expectation, accumulated clinical correlation over two centuries, or other mechanisms. This study does not adjudicate these possibilities; it evaluates the empirical alignment itself.

To our knowledge, only small-scale or manual approaches have attempted such validation. A 2018 study compared aluminum toxicology with homeopathic Alumina through literature review of a single remedy, finding 50.76% concordance without statistical significance testing [3]. Computational approaches using natural language processing have been applied to homeopathic repertorization (matching patient symptoms to remedies) [4,5], but not to validation of historical provings against modern clinical knowledge.

We hypothesized that embedding-based semantic similarity could detect meaningful alignment between historical proving symptoms and contemporary clinical signs across hundreds of remedies. We pre-specified quality-weighted statistics and significance thresholds to distinguish true signal from chance.

Materials and Methods

Study Design

This was a retrospective computational analysis comparing two text corpora: historical homeopathic symptom descriptions and modern clinical toxicology signs. No human subjects were involved.

Data Sources

Historical symptoms: We extracted 126,667 symptom descriptions from six digitized classical Materia Medica sources: Kent's Repertory (61,595), Hering's Guiding Symptoms (55,314), Boericke's Materia Medica (5,196), Allen's Encyclopedia (1,777), Lippe's Materia Medica (1,618), and Keynotes (1,167). These yielded 74,415 unique symptom texts across 862 remedies, of which 820 had associated clinical signs.

Modern clinical signs: A curated database of 4,091 clinical signs associated with toxic or pharmacological exposures was compiled using AI-assisted extraction (Claude, Anthropic) from contemporary toxicology and clinical pharmacology references. For each substance, the extraction prompt asked: "What are the clinical signs of [SUBSTANCE] toxicity or pharmacological exposure?"—framed in terms of the chemical substance identity (e.g., "atropine," "lead," "strychnine"), not the homeopathic remedy name. Prompts referenced toxicology and clinical pharmacology, not homeopathic literature. Each clinical sign includes a pharmacological rationale (e.g., "M3 receptor blockade in iris sphincter muscle" for mydriasis in atropine exposure). Extracted signs were validated against standard toxicology references for accuracy. Of the 4,091 signs, 4,003 (97.8%) were classified as "rare" (IDF ≥ 4.0), meaning they are specific to a small number of substances rather than generic findings.

Embedding Model

Text embeddings were generated using Qwen3-Embedding (0.6B parameters, 1024 dimensions), accessed via local deployment on Apple Silicon (RPlay inference server). Both historical symptom texts and modern clinical sign descriptions were embedded into the same vector space using mean pooling of token embeddings with L2 normalization. Random seed was fixed at 42 for reproducibility.

Matching Algorithm

For each remedy r with symptom set S_r and clinical sign set C_r:

1. For each clinical sign $c \in C_r$, compute cosine similarity to all symptoms
2. Record the maximum similarity score and matching symptom
3. Filter matches by similarity threshold (≥0.50) and IDF score (≥4.0, indicating "rare" signs)

Inverse document frequency (IDF) weighting downweighted common signs (e.g., “nausea”) that match many remedies non-specifically.

Test Statistics

We pre-specified one primary and three secondary endpoints:

Statistic	Type	Description
`top3_unique_weighted`	**Primary**	Σ(score × uniqueness) for top 3 remedy-specific matches
`top5_idf_weighted`	Secondary	Σ(score × IDF) for top 5 rare matches
`max_score`	Secondary	Maximum similarity among rare matches
`rare_count`	Secondary	Number of rare clinical signs with similarity ≥0.50

The primary endpoint was chosen because uniqueness-weighting emphasizes remedy-specific (pathognomonic) matches over generic findings.

Null Models and Permutation Testing

Two null models controlled for different confounds:

Null Model C (primary inferential model): Clinical signs fixed; symptoms randomly drawn from a size-matched remedy (quantile buckets). This tests whether the specific symptom corpus of a remedy aligns better than expected with its clinical signs.

Null Model A (robustness check): Symptoms fixed; clinical signs randomly drawn from a size-matched remedy. This controls for the possibility that certain symptom corpora generically match many clinical sign sets.

For each remedy, 2,000 permutations were performed under each null model. P-values were computed as (1 + k) / (1 + n), where k is the count of permutation statistics ≥ observed.

Multiple Testing Correction

Benjamini-Hochberg FDR correction was applied in two pre-specified ways:

1. **Full FDR:** All 820 remedies
2. **Gated FDR:** Pre-specified high-evidence subset meeting quality criteria:
3. Maximum similarity score ≥ 0.78
4. Unique rare matches ≥ 2
5. High-IDF matches (IDF ≥ 5) ≥ 2

The gated approach reduces multiple testing burden while focusing on remedies with high-quality matches. Gating thresholds were determined during initial method development using a small ex-

ploratory sample, then frozen before the full analysis was executed. This two-stage procedure follows standard practice for reducing multiple testing burden in high-dimensional analyses [see Sensitivity Analysis for threshold robustness assessment].

To account for testing four endpoints per remedy, we additionally report results at a Bonferroni-corrected threshold ($\alpha/4 = 0.0125$). This two-tier approach distinguishes remedies surviving the most conservative cross-endpoint correction from those significant at per-endpoint FDR only.

Sensitivity Analysis

To assess robustness, we evaluated stability of significant findings across: - Similarity thresholds: 0.50, 0.55, 0.60 - Gating max_score thresholds: 0.75, 0.78, 0.80 - IDF thresholds: 4.0, 4.5, 5.0

Enrichment Analysis

As an exploratory assessment, enrichment in the gated subset was evaluated using a one-tailed binomial test comparing observed significant remedies against the 5% expected under the null hypothesis of no true signal.

Reproducibility

Analysis code and data are available from the corresponding author upon reasonable request. The random seed (42) and model specifications ensure reproducible permutation results.

Results

Overall Statistics

Metric	Value
Remedies analyzed	820
Total symptoms (all sources)	126,667
Unique symptom texts	74,415
Clinical signs	4,091
Remedies passing pre-specified gate	78 (9.5%)

Primary Endpoint: top3_unique_weighted

Under gated FDR correction (n=78), **14 remedies (17.9%)** achieved significance at α=0.05 on the primary endpoint.

Under full FDR correction (n=820), **41 remedies (5.0%)** showed nominal significance, though this should be interpreted cautiously given the high multiple testing burden.

Secondary Endpoints

Statistic	Gated FDR (n=78) p<0.05	Full FDR (n=820) p<0.05
`top3_unique_weighted` (primary)	14	41
`top5_idf_weighted`	20	33
`max_score`	5	18
`rare_count`	0	0

Combining across all four endpoints, **25 of 78 gated remedies (32.1%)** achieved significance on at least one endpoint at per-endpoint FDR (α = 0.05). Of these, **22 remedies (28.2%)** survived Bonferroni correction for multiple endpoints (α/4 = 0.0125). The remaining 3 remedies were significant at per-endpoint FDR only.

Negative Controls

To verify method specificity, we examined three substances expected to show no toxicological signal: Saccharum lactis (lactose, the classic homeopathic placebo vehicle), Alcohol (ethanol, a com-

mon solvent), and Teplitz aqua (mineral spring water). None passed the pre-specified quality gate. All raw p-values exceeded 0.14 across all four endpoints.

Table 2. Negative Control Substances

Substance	Type	Symptoms	Clinical Signs	Min p (any endpoint)	Badge
Saccharum lactis	Placebo vehicle (lactose)	47	7	0.14	NONE
Alcohol	Solvent (ethanol)	6	36	1.00	NONE
Teplitz aqua	Mineral water	183	10	0.18	NONE

Inert substances showed no semantic alignment signal, confirming that the method does not generate false positives from pharmacologically inactive materials.

Enrichment (Exploratory)

Per-endpoint FDR tier (25 of 78): - Expected significant by chance: 78 × 0.05 = 3.9 remedies - Observed significant (any endpoint): 25 remedies - Enrichment ratio: 6.4-fold - Binomial test (exploratory): $p = 3.8 \times 10^{-14}$ (one-tailed)

Bonferroni-corrected tier (22 of 78): - Expected significant by chance: 78 × 0.0125 ≈ 1.0 remedy - Observed significant (any endpoint at α/4): 22 remedies - Enrichment ratio: 22.6-fold - Binomial test (exploratory): $p = 9.8 \times 10^{-24}$ (one-tailed)

Both enrichment estimates should be interpreted as hypothesis-generating given the gated design. The Bonferroni-corrected tier shows stronger enrichment, indicating that the most robust findings concentrate in a smaller subset.

Clinically Interpretable Alignments

Table 1 presents a selection of remedies achieving gated FDR significance with clinically recognizable matches.

Table 1. Selected Remedies with Significant Semantic Alignment

Remedy	Substance	Top Clinical Match	Similarity	p (gated FDR)	Tier
Belladonna	Atropine alkaloid	Dysphagia	0.918	0.036	FDR
Antimonium tart.	Antimony	Thick white tongue coating	0.901	0.003	Bonf.
Chelidonium	Chelidonine alkaloid	Right subscapular pain pattern	0.887	0.003	Bonf.
Thuja	Thujone	Anogenital lesions	0.846	0.008	Bonf.
Strychninum	Strychnine	Risus sardonicus	0.782	0.003	Bonf.
Plumbum met.	Lead	Burton's line (lead line)	0.796	0.004	Bonf.
Sepia	Cuttlefish ink	Chloasma/melasma pattern	0.782	0.003	Bonf.
Mygale	Spider venom	Chorea-like movements	0.826	0.003	Bonf.
Lachesis	Snake venom	Gingival bleeding	0.793	0.003	Bonf.
Kali iodatum	Potassium iodide	Coryza (iodide rhinitis)	0.798	0.003	Bonf.

These matches are consistent with known toxidromes. Belladonna's alignment with dysphagia is consistent with anticholinergic toxidrome. Plumbum (lead) matched Burton's line (the characteristic blue gum line of lead poisoning, similarity 0.796) as well as wrist-related weakness (0.720). Strychninum's match to risus sardonicus (the sardonic grin of strychnine poisoning) is pathognomonic. Chelidonium's right subscapular pain pattern corresponds to hepatobiliary toxicity of chelidonine alkaloids.

Sensitivity Analysis

Results were stable across gating threshold variations:

Gating max_score	Remedies passing	Significant (any endpoint)	% Significant	Enrichment	Binomial p
0.75	136	31	22.8%	4.6×	1.1×10^{-12}
0.78 (pre-specified)	78	25	32.1%	6.4×	3.8×10^{-14}
0.80	55	16	29.1%	5.8×	7.0×10^{-9}

The enrichment ratio remained between 4.6× and 6.4× across all thresholds, with binomial p-values consistently below 10^{-8}. The proportion of significant remedies was stable (23–32%), suggesting the finding is not sensitive to the choice of gating threshold.

Null Model Comparison

Results were consistent across both null models. For significant remedies, Null Model A (BioAI shuffle) generally showed similar or slightly higher p-values, supporting robustness of findings.

Discussion

Principal Findings

This comprehensive computational analysis reveals statistically significant semantic alignment between a subset of 19th-century homeopathic provings and modern clinical toxicology. Among pre-specified high-evidence remedies, observed significance rates substantially exceeded chance expectations.

Interpreting the Proportion of Significant Remedies

That 25 of 820 remedies (3.0%) achieved per-endpoint FDR significance—of which 22 survive Bonferroni correction for multiple endpoints—warrants careful interpretation. This proportion does not indicate weakness of signal—rather, it reflects the expected distribution of detectable toxidromes across the remedy corpus. The majority of remedies are derived from substances without distinctive, pathognomonic clinical presentations (e.g., mineral salts, common plant extracts). For these, no strong alignment with specific toxicological signs would be expected even under the hypothesis that provings reflected substance-specific physiological patterns.

The statistical evidence rests not on the absolute count of significant remedies, but on the enrichment: among remedies meeting quality criteria, 32.1% were significant at per-endpoint FDR versus the 5% expected by chance—a 6.4-fold excess ($p = 3.8 \times 10^{-14}$). Under the more stringent Bonferroni correction, 28.2% remained significant (22.6-fold enrichment, $p = 9.8 \times 10^{-24}$). Moreover, the significant remedies are not randomly distributed: they correspond precisely to substances with well-characterized toxidromes (lead poisoning, anticholinergic syndrome, strychnine toxicity, hemlock paralysis). This convergence between statistical significance and clinical plausibility strengthens confidence in the finding.

Clinical Interpretation

The aligned matches are clinically specific. Plumbum (lead) matched both Burton's line (the characteristic blue gum line, similarity 0.796) and extensor weakness (0.720)—hallmarks of chronic lead poisoning. Belladonna provers noted difficulty swallowing—a cardinal feature of anticholinergic syndrome. Strychninum matched risus sardonicus, the pathognomonic facial spasm of strychnine poisoning. These correspondences suggest that some historical proving observations may reflect recognizable toxicological patterns.

Importantly, simple match counts (`rare_count`) failed to detect signal in either gated or full FDR analyses; quality-weighted statistics emphasizing high-similarity and remedy-specific matches were necessary. This indicates that discriminating signal requires attention to match quality, not merely quantity.

Comparison with Prior Work

We found only small-scale or manual prior attempts at proving validation. The 2018 Alumina study [3] compared a single remedy through literature review, finding ~50% concordance without statistical testing. Computational NLP approaches in homeopathy have focused on repertorization [4,5] rather than historical validation. Our approach scales to hundreds of remedies with rigorous null model comparison.

Limitations

1. **Semantic vs. clinical validity:** Embedding similarity captures linguistic/semantic alignment, which may include superficial textual overlap rather than true clinical correspondence. Expert adjudication of top matches would strengthen clinical interpretation.

2. **Gated analysis:** While gating criteria were pre-specified based on pilot exploration, reviewers may consider this approach as introducing selection. We emphasize that the gated analysis supports an enrichment claim (more significant remedies than expected by chance), not a discovery claim about individual remedies. The sensitivity analysis showing consistent results across thresholds, and the ungated full-FDR results showing no inflation, partially mitigate this concern.

3. **Potential for LLM-mediated circularity:** Clinical signs were extracted using a large language model, which may have been trained on texts containing both toxicological and homeopathic information. Although prompts were framed exclusively in terms of substance toxicology (e.g., "clinical signs of lead poisoning"), not homeopathic remedy profiles, we cannot fully exclude the possibility that the model's training data created indirect overlap between the two corpora. To illustrate our extraction approach: for Plumbum metallicum (lead), the prompt asked for clinical signs of lead poisoning—producing "Burton's line on gingival margin" with the rationale "lead sulfide deposition at the gum–tooth junction." This is standard clinical toxicology, verifiable in any toxicology reference (e.g., Goldfrank's Toxicologic Emergencies), and makes no reference to homeopathic proving literature. Each of the 4,091 clinical signs was generated and validated in this manner. This concern is further mitigated by the negative control results (inert substances produced no signal) and by the fact that pathognomonic signs like Burton's line or risus sardonicus are well-established in clinical toxicology independently of any homeopathic literature. Future work could strengthen this defense through comparison with manually curated toxicology databases (e.g., CDC toxidrome tables, Goldfrank's).

4. **Embedding model specificity:** Qwen3-Embedding (0.6B) is a general-purpose text embedding model, not specifically trained or validated on medical text. Its selection was driven by local deployment capability and computational efficiency. Replication with domain-specific models (e.g., BioBERT, PubMedBERT) or larger general-purpose models (OpenAI text-embedding-3-small) would strengthen generalizability. Additionally, embedding similarity is sensitive to text length: detailed multi-clause symptom descriptions may score lower than terse rubrics against short clinical sign descriptions.

5. **Therapeutic implications:** These findings do not support homeopathic treatment efficacy. This study is agnostic to mechanism and evaluates only informational coherence between historical texts and modern clinical knowledge. Validation of historical symptom observations is independent of debates about ultra-high dilution mechanisms or clinical outcomes.

Why Only Toxic Substances Show Signal

The concentration of significant findings among substances with well-characterized toxidromes is expected, not surprising. Substances like lead, strychnine, and belladonna produce distinctive, organ-specific clinical presentations (Burton's line, risus sardonicus, anticholinergic syndrome) that generate unique textual signatures in both historical provings and modern toxicology references. By contrast, substances without distinctive toxidromes—common minerals, mild plant extracts, nosodes—lack the pathognomonic specificity needed to produce above-chance semantic alignment. This selectivity is itself evidence of method validity: a method that found signal everywhere would be suspect.

Implications

Signal concentrates in remedies with pathognomonic signs, suggesting that some proving descriptions align with recognized toxicological patterns for substances with distinctive clinical presentations. This framework could be generalized to other historical medical corpora with modern clinical anchors.

Conclusions

This comprehensive analysis demonstrates statistically significant semantic alignment between a subset of 19th-century homeopathic provings and modern clinical toxicology. Among remedies meeting pre-specified quality criteria, alignment rates substantially exceeded chance expectations (22 remedies surviving Bonferroni correction for multiple endpoints, 22.6-fold enrichment, binomial $p = 9.8 \times 10^{-24}$). Negative control substances (lactose, ethanol, mineral water) showed no signal, confirming method specificity.

These findings support the presence of clinically recognizable toxicological patterns in some historical proving descriptions, independent of any claims regarding homeopathic therapeutic efficacy. Simple match counts fail to detect this signal; quality-weighted statistics emphasizing uniqueness and high similarity are required.

Future work should include expert adjudication of top matches, replication with alternative embedding models, extension to non-toxic remedy categories, and separate analysis of provings with explicit potency/dose metadata to disentangle material-dose from ultra-high-dilution observations.

References

1. Hahnemann S. Organon of Medicine. 6th ed. 1842.
2. Ernst E. A systematic review of systematic reviews of homeopathy. Br J Clin Pharmacol. 2002;54(6):577-82.
3. Teut M. Comparison of aluminum toxicology and homeopathic Alumina: a mixed methods study. Int J High Dilution Res. 2018;17(3-4):897.
4. Krishnamoorthy R. Using Similarity Search in Homeopathy Repertorization Software. 2024. Available: https://www.rangakrish.com/
5. Singh SR, Patil AD. Natural Language Processing as an AI Tool in Prognostic Factor Research in Homeopathy. Indian J Integr Med. 2024.
6. Boericke W. Pocket Manual of Homoeopathic Materia Medica. 9th ed. 1927.
7. Kent JT. Repertory of the Homoeopathic Materia Medica. 1897.
8. Hering C. The Guiding Symptoms of Our Materia Medica. 10 vols. 1879-1891.
9. Allen TF. The Encyclopedia of Pure Materia Medica. 10 vols. 1874-1879.

Supporting Information

S1 Table. Complete results for all 820 remedies analyzed. **S2 Table.** Sensitivity analysis across threshold variations. **S1 Code.** Analysis scripts and reproducibility materials. **S1 Data.** Deduplicated symptom corpus and clinical sign database.

Acknowledgments

The author is not affiliated with any pharmaceutical or homeopathic company. This work was conducted as independent computational research.

Author Contributions

Single-author paper. GP conceived and designed the study, developed the software, performed the analysis, and wrote the manuscript.

Funding

No external funding was received for this work.

Competing Interests

The author declares no competing interests. The author is an independent researcher with no affiliations to pharmaceutical or homeopathic organizations.

Data Availability

Analysis code and data are available from the corresponding author upon reasonable request. Random seed (42) and model specifications enable full reproducibility.