Historical Medical Knowledge Graphs and Ontologies from the Medical History of British India Corpus (1850–1950)

Mehrdad Almasi, Tugce Karatas
Luxembourg Centre for Contemporary and Digital History, University of Luxembourg, Luxembourg
{mehrdad.almasi, tugce.karatas}@uni.lu

Abstract

This research presents a reproducible framework for constructing biomedical knowledge graphs and ontologies from digitized historical archives. Focusing on the *Medical History of British India* corpus (468 reports; ∼22.5M words; 1850–1950), our pipeline combines BioBERT-based entity recognition, LLM-guided relation extraction with LLM-based filtering, and clustering-based ontology induction. Reliability is strengthened through canonicalization, schema mapping to standardized biomedical relation types, and multi-metric edge scoring with temporal decay; a manual evaluation of 500 validated triples yields 0.892 precision. The resulting resources comprise $282{,}882$ extracted relations, consolidated into $22{,}360$ unique surface forms and organized into 71 thematic clusters. Frequent categories include *After Treatment* (∼1,242 mentions), *Date of Inoculation* (∼540), and diverse causal relations, while the induced ontology highlights six epidemic diseases: plague, cholera, malaria, kala-azar, leprosy, and smallpox together with their characteristic interventions (e.g., quinine therapy, vaccination campaigns, hospital disinfection). Temporal analyses capture historically plausible trajectories: plague interventions peaking in the 1890s, cholera's long-run decline, and tuberculosis departments rising after 1910. All code, relation inventories, ontologies, and visualizations are released in a GitHub Repository, enabling reproducibility and supporting research in historical NLP, biomedical informatics, and digital humanities.

Keywords: Historical NLP, Knowledge Graph Induction, Biomedical Ontologies, Digital Humanities

1. Introduction

The study of medical history offers insights into how societies have understood, managed, and responded to disease across time. Historical records not only document medical breakthroughs and the emergence of public health infrastructures but also reveal the interplay of medicine, governance, and community responses. By systematically analyzing such archives, researchers can trace the evolution of medical traditions and practices, identify persistent challenges in disease management, and better understand the roots of present-day public health systems (Mishra and Shridevi, 2024). In particular, digitized records from large-scale repositories allow us to revisit long spans of medical history with new analytical tools, making patterns of knowledge production and dissemination more visible than ever before.

Knowledge graphs (KG) provide a structured approach to representing and analyzing complex historical data. They enable the integration of heterogeneous information such as diseases, treatments, institutions, and locations into networks that can be queried, visualized, and compared across time. Recent advances in natural language processing and graph representation learning have made it possible to construct large, semantically meaningful knowledge graphs directly from text (Peng et al., 2023; Choi and Jung, 2025; Yang et al., 2023). Biomedical ontologies such as UMLS (Bodenreider, 2004), MeSH (Lipscomb, 2000), and SNOMED CT (Stearns et al., 2001) have played a central role in standardizing medical vocabularies and relations, supporting interoperability across biomedical databases and clinical systems. When applied to historical corpora, aligning extracted entities and relations with these ontologies not only improves reliability but also allows direct comparison between past medical practices and present biomedical knowledge. Applied to historical medical sources, this approach facilitates both macro-level overviews (e.g., the rise and decline of particular diseases) and micro-level case studies (e.g., the spread of a treatment method). By bridging text mining with historical scholarship, knowledge graphs enable interdisciplinary research and the large-scale study of medical history.

Biomedical ontology and KG resources provide a foundation for normalizing noisy extractions and comparing signals across corpora. The *Unified Medical Language System (UMLS)* (Bodenreider, 2004) offers a meta-thesaurus and semantic network that enable cross-vocabulary linking of diseases, drugs, and procedures, while *SemMedDB* (Kilicoglu et al., 2012) aggregates subject–predicate–object "semantic predications" mined from PubMed at scale. Our pipeline leverages this approach by mapping induced clusters to standardized biomedical relation types and provenance-aware aggregation to handle historical spellings, OCR artifacts, and administrative discourse (Lindberg et al., 1993; Kilicoglu et al., 2012).

Our approach requires the identification of named entities, for which we employ *BioBERT*. This model enhances biomedical NER and rela-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7781–7790
11–16 May 2026.

tion extraction through pretraining on PubMed and PMC corpora (Lee et al., 2020). However, historical sources such as newspapers and medical reports introduce additional challenges, including orthographic variation, genre differences, and OCR noise. These issues have been systematically studied in the HIPE (Historical and Multilingual Information Processing for Named Entity Recognition) shared tasks, which provide benchmarks for NER and entity linking in noisy historical texts (Ehrmann et al., 2020). To connect fine-grained extractions with broader historical themes, we further incorporate *BERTopic*, which clusters transformer-based embeddings and produces interpretable topic labels that align with our relation inventories (Grootendorst, 2022). The combination of BioBERT for biomedical mentions , LLM-guided triple extraction (named entity, relation, named entity), and BERTopic for thematic alignment enables us to recover both canonical disease–intervention links and institutional practices across one century.

Building on this perspective, our work focuses on the *Medical History of British India* corpus, a large-scale digitized collection of official medical and public health publications spanning 1850–1950. This century-long archive documents governmental responses to epidemic diseases such as plague, cholera, malaria, kala-azar, leprosy, and smallpox, alongside the gradual shift from humoral to bacteriological frameworks of medicine. It also records the establishment of hospitals, vaccination programs, and veterinary departments, providing a unique window into the evolution of public health infrastructures during this period. Despite its richness, the corpus presents significant challenges for computational analysis, including OCR artifacts from scanned images, historical spellings and terminology, and the intermixing of medical, administrative, and socio-cultural discourse.

To address these challenges, we develop a pipeline for automatically inducing biomedical knowledge graphs from this corpus. Our approach combines domain-specific embeddings, clustering-based ontology induction, and large language model (LLM) filtering to extract reliable entities and relations from noisy text at scale. We further introduce canonicalization of surface forms, schema mapping to standardized biomedical relation types, and multi-metric edge scoring (lift, NPMI, jacc_doc, time_decayed_support, frequency, cohesion) to enhance robustness and interpretability. The resulting historical biomedical ontologies and relation inventories make visible both the principal disease burdens documented in the corpus and the interventions deployed against them (e.g., quinine therapy, vaccination campaigns, and disinfection practices), while also surfacing institutional perspectives (e.g., hospital admissions, tuberculosis departments, animal-based vaccination experiments).

Contributions.

- To our knowledge, the first induced biomedical ontology constructed from historical colonial medical reports (1850–1950).
- Integration of temporal decay and topic alignment into ontology induction, rarely applied in historical NLP.
- Public release of scored edges, relation inventories, and ontologies (not only code) enabling reproducibility.
- A transparent pipeline (NER, LLM validation, canonicalization, multi-metric scoring) explicitly designed to cope with OCR noise and historical variation.

Paper structure. Section 2 describes the corpus and details the induction pipeline and scoring. Section 3 reports ontology quality, topic–relation alignment, and historical case studies. Section 4 concludes with limitations and future directions.

Statistic	Value	Notes
Documents	468	Unique medical reports
Pages	117,022	OCR-processed pages
Words	∼22.5M	OCR-extracted tokens
Images	120,903	Scanned page images
Time span	1850–1950	Century of medicine
Main diseases	6 categories	Cholera, Plague, Malaria, Leprosy, Smallpox, Kala-azar (metadata)
Source	NLS	National Library of Scotland archive
Availability	Hugging Face	Publicly accessible

Table 1: Key statistics of the *Medical History of British India* dataset (metadata values).

2. Methodology

2.1. Corpus Description

The *Medical History of British India* corpus contains 468 digitized medical and public health publications produced between 1850 and 1950. It comprises ∼22.5M words across 117,022 OCR-processed

pages, with 120,903 scanned images and associated metadata (Table 1). Documents include annual health reports, epidemic investigations, disease histories, vaccination statistics, and veterinary records. The dataset used in this study is the *Medical History of British India* collection (of Scotland, 2019), openly available via Hugging Face. All OCR text is sourced from the National Library of Scotland's digitized corpus. We do not benchmark OCR accuracy, as this lies outside the scope of the present study. Instead, our pipeline is explicitly designed to compensate for OCR noise through LLM-based filtering, canonicalization, and multi-metric reliability scoring. [1]

2.2. Pipeline Overview

We implement a multi-stage pipeline for inducing biomedical knowledge graphs and ontology from the *Medical History of British India* corpus. The design combines biomedical NER, LLM-based relation extraction, clustering, schema induction, and multi-metric reliability scoring, with all processing performed in a streaming fashion to handle both the size and noise of the corpus (Figure 1).

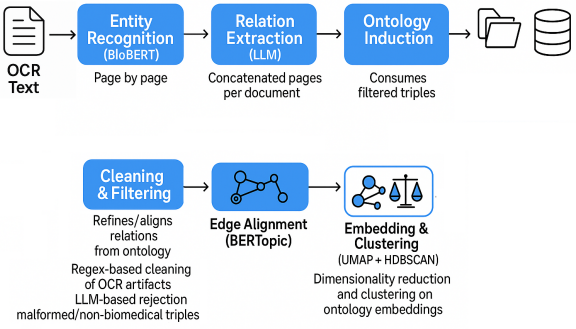

Figure 1: Pipeline for inducing biomedical knowledge graphs from medical history text. The top row shows the main extraction workflow (OCR, entity recognition, relation extraction, ontology induction, outputs), while the bottom row expands key components in more detail (cleaning and validation, edge alignment, and embedding-based clustering).

The process begins with entity recognition, where biomedical mentions are detected on a per-page basis using BioBERT (`d4data/biomedical-ner-all`). In the next step, candidate triples consisting of subject, relation, and object are generated by an instruction-tuned large language model (`qwen2.5:32b-instruct`, accessed via Ollama). To improve robustness, relation extraction is performed on concatenated document text, and results are written directly to JSONL and CSV files. Cleaning and filtering steps remove OCR artifacts through regex-based rules and reject malformed or non-biomedical triples through LLM-based validation.

Entity and relation strings are then embedded with BioBERT and clustered using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes et al., 2017), producing canonical forms that mitigate spelling variability and OCR-induced noise. Clusters are mapped into schema-level biomedical relation types (e.g., treated_with, causes, after_treatment, vaccinated_against, located_in), yielding an induced ontology in JSON and tabular form. Vaccination and inoculation variants are explicitly grouped under `vaccinated_against` to preserve historically meaningful terminology.

Relations are aggregated across documents and weighted by a combination of frequency, lift, normalized pointwise mutual information (NPMI), Jaccard similarity, time-decayed support (based on publication year), and embedding-based cohesion. Finally, BERTopic is applied to document-level representations, allowing relation clusters to be aligned with thematic topics, which enables analyses of how diseases, treatments, and institutions were discussed across time and context.

Filtering and LLM validation. Lightweight filters remove boilerplate (stoplisted relations, trivial arguments, and near-duplicates) before scoring, while LLM validation ensures biomedical specificity.

For LLM-based validation, we prompt a local instruction model (Ollama) to check whether a candidate triple is explicitly supported in context, returning only `ACCEPT` or `REJECT`. Results are cached in `llm_filter_cache.jsonl`, so identical triples are never re-queried; in total, ∼282k raw triples were filtered down to ∼104k validated relations. The surviving triples are written to CSV/JSONL, forming the input for ontology induction and scoring. These lightweight filters suppress boilerplate artifacts, while LLM validation ensures biomedical specificity.

2.3. Reliability Measures

Reliability is reinforced through several mechanisms. Canonicalization yields stable identifiers for entities and relations by consolidating spelling variants and reducing redundancy. Large language model (LLM) filtering suppresses OCR artifacts and non-biomedical triples, while multi-metric edge scoring integrates statistical association measures with temporal weighting and

[1] https://huggingface.co/datasets/davanstrien/MedicalHistoryofBritishIndia

embedding-based cohesion to prioritize relations that are both frequent and semantically meaningful. We combine statistical association, temporal weighting, and embedding cohesion (columns: lift, NPMI, `jacc_doc`, `timedecayed support`, `s_cohesion`, `r_cohesion`, `o_cohesion`, and a final score). `jacc_doc`(`jacc_doc`) is computed between the set of documents supporting an edge and the union of its subject's and object's document sets, capturing how concentrated an edge is relative to its incident (vertex) nodes. Temporal weighting uses an exponential half-life of 15 years relative to the corpus median year: $w(y) = 0.5^{|y-y_{\text{med}}|/15}$, and `time-decayed support` is the count weighted by $w(y)$. For headline counts, we aggregate many lexical variants into canonical relation labels, and report totals *after* canonicalization. The resulting canonical labels and schema assignments are released in `relation_inventory.csv` and `ontology_clean.json`, which enable exact reproduction of the reported frequencies for categories such as *After Treatment*, *Cause/Caused By*, and *Date of Inoculation*.

2.4. Outputs

The pipeline outputs three main resources. First, ontology inventories that record the canonical entities, relations, and schema mappings, released in both raw and cleaned JSON files (`ontology.json`, `ontology_clean.json`) alongside tabular inventories (`relation_inventory.csv`, `relation_types.csv`, `nodes.csv`). Second, scored graph exports in CSV and GraphML format (`edges_aggregated.csv`, `graph.graphml`), supporting visualization and integration with network analysis tools. Third, topic–relation mappings are released separately after alignment with BERTopic, as CSV files (`ontology_topics.csv`, `relation_topics.csv`, `relation_topics_all.csv`, `topic_top_relations.csv`) and visualizations (`ontology_topics_network` in html, png, pdf), enabling both macro-level trend analysis and micro-level case studies.

This separation reflects the pipeline design: relation extraction and ontology induction produce canonicalized inventories, while topic alignment (run as a downstream stage) supplies conservative surface-string counts and thematic groupings (GitHub Repository).

To document the released schema, we provide a schema-level Entity–Relationship (ER) overview in Figure 2, summarizing the core entity types and mapped biomedical relation types used to structure the knowledge graph. For readability, the diagram omits long-tail relation phrases that are not mapped to our schema.

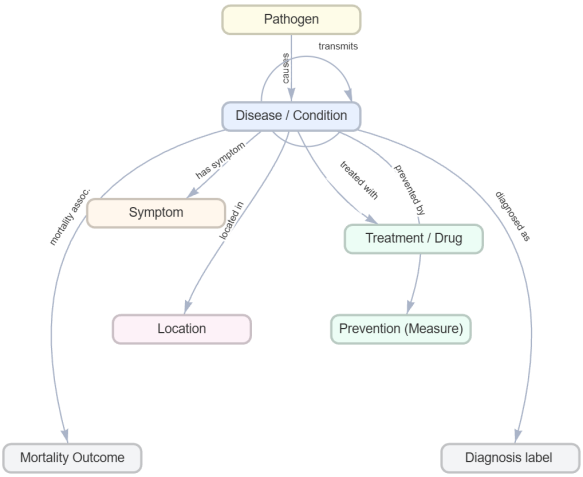

Figure 2: Schema-level ER diagram of the induced historical medical ontology (filtered). Boxes denote core entity types; arrows denote mapped schema relations (e.g., *causes*, *treated_with*, *has_symptom*, *located_in*). Unmapped/long-tail relation phrases are omitted for clarity.

3. Results

Our results highlight both the reliability of the induced ontology and the historical insights it enables. We present findings on ontology induction, relation inventories with temporal trends, topic–relation alignments, and illustrative case studies. Together, these demonstrate how noisy medical records can be transformed into structured resources for large-scale historical and biomedical analysis.

3.1. Ontology Induction and Relation Analysis

The induced graph comprises 43,208 nodes, 42,621 raw relation strings, and 168,030 aggregated edges, derived from biomedical relations extracted across 468 corpus volumes (HuggingFace dataset). Due to identifier splits in the extraction logs, relation files enumerate 469 document instances, but these correspond to 468 unique parent volumes in the metadata. Edge weights are long-tailed: 91.6% of edges occur only once (max. 128). Despite sparsity (median Jaccard = 0.0007), high-lift edges are reliable (median = 10.84; 95th percentile $\approx$ 366). Self-loops account for 7.03% (11,810 / 168,030), consistent with segmentation artifacts; canonicalization reduces duplicates but does not eliminate all within-entity mentions. These diagnostics highlight both the challenges of sparsity and the robustness of multi-metric scoring.

Ontology induction surfaces a broad biomedical and institutional vocabulary. BERTopic yields

71 thematic clusters; the relation inventory contains **22,360** unique cleaned relation surface forms. Across the corpus, these correspond to **104,269** normalized (canonicalized) triple instances (95,024 under conservative matching). High-frequency categories include *After Treatment* (2,475 mentions), *Date of Inoculation* (1,099), and causal families (4,471 overall). Distributions under canonical vs. conservative counting are consistent, reflecting lens rather than inconsistency. Cross-checks against UMLS, MeSH, and SNOMED confirm alignment with attested biomedical relations. Manual spot-checks (e.g., plague→hospital admission, smallpox→vaccination) further reinforce substantive validity.

Across schema-level aggregation, the six principal epidemic diseases dominate: smallpox (278 mentions), plague (308), cholera (215), malaria (144), leprosy (102), and kala-azar (10). For completeness, we also track tuberculosis (45), which—while not part of the six principal epidemic foci—appears in institutional contexts and is analyzed for temporal trends below. Canonical labels (e.g., Smallpox, Pox, Variola; Leprosy, Leper) ensure consistency. Tables 4 and 5 summarize ontology components and themes.

3.2. Evaluation

Because exhaustive gold-standard annotation is infeasible at the scale and noise level of century-spanning OCR archives, we combine (i) a small manual estimate of extraction quality, (ii) intrinsic statistical diagnostics, (iii) ablation analyses, and (iv) historical plausibility checks. Together, these provide complementary evidence for reliability under realistic constraints.

3.2.1. Manual Precision and Recall Probe

To obtain a direct quality estimate, we manually evaluated a high-confidence subset of extracted triples. We sampled 5 batches of 100 validated triples (n=500) and checked each triple against its page-level OCR context. A triple was marked *correct* if the asserted relation was explicitly supported in context, allowing for minor OCR and spelling variation after canonicalization; otherwise it was marked *incorrect*. This yields a precision of 0.892 (446/500), with batch-level precision ranging from 0.88 to 0.91.

Estimating recall is harder in an open-world setting, because the set of potentially valid relations on a page is not closed and depends on interpretation. As a lightweight probe, we manually listed additional plausible relation statements on a small set of pages and compared them against extracted triples using relaxed (fuzzy) string matching over canonicalized forms. On average, we identified roughly 20 candidate relations per 100 that were not recovered by the pipeline, suggesting an approximate recall on the order of 0.83 under this relaxed matching. We report this only as an indicative lower bound, since relation boundaries in historical prose can be subjective.

3.2.2. Intrinsic Diagnostics

The induced graph is highly sparse: the median Jaccard similarity between edge-supporting documents is only 0.0007 (based on edge diagnostics in `edge_diagnostics.json`). Such sparsity is expected in century-spanning historical corpora, where most disease–intervention relations appear only once or in a temporally narrow cluster. At the same time, high-lift edges remain reliable: the distribution of *lift* is long-tailed (median = 10.84; 95th percentile ≈ 366), while *NPMI* and composite *scores* reveal a sharp contrast between the bulk of low-weight associations and a small subset of very strong signals. Rather than indicating weakness, this skewness underscores the value of our multi-metric scoring design: a single metric alone would either over-emphasize rare pairs (lift) or underweight historically bursty associations (frequency). Self-loops account for $\approx 7\%$ (11,810 of 168,030 edges), consistent with segmentation artifacts; canonicalization reduces duplicates but does not eliminate all within-entity mentions. Taken together, these diagnostics highlight both the challenges of sparsity and the robustness of the scoring framework.

3.2.3. Ablation Analyses

We conducted two ablations to isolate the contributions of temporal weighting and LLM-based validation.

Ranking stability. We compared edge rankings produced with vs. without temporal decay, and against association-only and cohesion-only baselines. Stability was measured using `Jaccard@K` overlap between ranked edge sets. At small cutoffs, Jaccard values near zero show that temporal weighting substantially reprioritizes the very top-ranked edges, surfacing historically localized relations such as plague→hospital admission during epidemic crises. At larger cutoffs, stability increases (Jaccard@500 > 0.608), demonstrating that although temporal decay alters the leading edges, the overall ontology remains globally robust. This confirms that temporal weighting enhances historical interpretability without destabilizing the graph as a whole. **Note.** Jaccard@K measures the set overlap between the top K ranked edges under two scoring schemes. A value of 0 indicates

completely different top-K sets, while a value of 1 indicates identical sets.

LLM filtering. We compared raw extractions against validated relations. Raw outputs included many administrative or boilerplate relations (e.g., treasury or establishment accounting). After LLM validation, the number of distinct relations decreased, the "admin lexicon" leakage ratio dropped, and the normalized count of unique biomedical relations per 1,000 edges also declined. This shift provides an indirect precision signal: fewer noisy relations and more consistent biomedical triples. Thus, validation demonstrably suppresses OCR-induced artifacts and non-biomedical material.

3.2.4. Historical Plausibility

Finally, we assessed whether high-scoring pairs aligned with historically documented medical practices. We extracted the top disease–intervention pairs overall and by era (pre-1900, 1900s, 1910s, 1920+; Table 2). The results reveal **plausible temporal trajectories**:

- Smallpox→vaccination dominates the 19th century and persists across all eras, reflecting sustained inoculation campaigns.
- Plague→hospital admission/disinfection peaks in the 1900s, coinciding with major plague epidemics, and then declines.
- Tuberculosis→institutional care rises after 1910, consistent with the expansion of tuberculosis departments and sanatoria.
- Malaria→quinine therapy/vector control shows continuity across eras, reflecting both long-term treatment and colonial public health policy.

These trajectories are consistent with secondary scholarship on colonial medicine and demonstrate that the induced graphs capture not only lexical associations but also the shifting epidemiological priorities of the state.

Our evaluation demonstrates that: (i) intrinsic metrics capture both sparsity and reliable high-weight signals; (ii) ablation analyses confirm the interpretive gains of temporal decay and the precision improvements of LLM filtering; and (iii) historical plausibility checks recover trajectories that align with attested practices. Taken together, these results provide strong evidence that the induced ontologies are reliable, interpretable, and valuable for historical biomedical research.

Era	Top disease–intervention pairs
Overall	Smallpox → Vaccination; Plague → Hospital admission; Malaria → Quinine therapy
Pre-1900	Smallpox → Vaccination; Cholera → Sanitation; Leprosy → Asylum care
1900s	Plague → Hospital admission/disinfection; Smallpox → Vaccination; Cholera → Water-supply measures
1910s	Tuberculosis → Institutional care; Smallpox → Vaccination; Malaria → Quinine therapy
1920+	Malaria → Quinine/vector control; Tuberculosis → Sanatorium treatment; Leprosy → Departmental administration

Table 2: Top-ranked disease–intervention pairs overall and by era, derived from high-scoring edges in `top_pairs_*.csv`. The pairs correspond to characteristic medical responses documented in colonial reports.

3.3. Topic Alignment

BERTopic alignment reveals how biomedical relations intersect with thematic discourses. The diseases-related relationships are clustered with topics of sanitation, hospitals, and disinfection, reflecting the management of crisis-driven epidemics. Malaria relations align with quinine therapy, vector control, and seasonal cycles, highlighting long-term therapeutic continuity. Smallpox relations are associated with vaccination campaigns and geographic coverage reports, illustrating the expansion of preventive medicine. Crucially, alignment also surfaces non-disease topics such as ganja and bhang consumption, showing that medicine extended beyond biomedicine to encompass intoxicants, nutrition, and social practices. Table 3 lists representative high-frequency surface-form relations for selected topics, as directly observed in the extraction inventory (`relation_topics.csv`); topic labels are automatically suggested by BERTopic and manually shortened for readability (very long relation strings are truncated for layout clarity; full versions are available in the released CSV files). These illustrate the raw lexical diversity that our pipeline subsequently consolidates into canonical categories (Sec. 3.1 and Tables 4–5). Topic–relation mappings are released as CSV files to enable reproducible exploration.

Relation extraction yields a large inventory of biomedical links, with multi-metric scoring ensuring reliability. Using *conservative normalized strings* (no schema aggregation), we compute counts by summing the `count` field over exact surface forms in `relation_topics.csv`. Under this lens, the *causal family* (surface strings contain-

ing "cause") totals **2,413** mentions (**2.54**%) out of $\mathbf{N}{=}95{,}024$ extractions; *After Treatment* occurs **1,242** times (**1.31**%), and *Date of Inoculation* **534** times (**0.56**%).[2] These rankings hold across topic thresholds and illustrate how frequent administrative/causal formulations coexist with more specific clinical reporting. (For canonicalized, schema-level tallies aggregated over lexical variants, see `relation_inventory.csv`.)

3.4. Case Studies

To illustrate the interpretive potential of the induced graphs, we present two concise case studies linking qualitative signals to the quantitative inventories reported above.

- **Plague hospitals.** Relations concentrate on disinfection, admissions, and relapse management, reflecting how containment blended biomedical and administrative measures. This aligns with ontology salience (308 mentions across 7 clusters) and topical structure (Table 3).
- **Smallpox vaccination.** Relations link vaccination to campaigns and geographic coverage. Smallpox (often surfaced as Pox) registers 278 mentions across 2 clusters and 9 forms. Inoculation and treatment appear frequently within vaccination- and hospital-adjacent topics.

These cases highlight how the graphs capture both practices (disinfection, inoculation) and infrastructures (plague hospitals, tuberculosis departments), coherently linking topic structure (Table 3) with ontology inventories.

3.5. Historical Interpretation

We project topic–relation evidence into a disease→intervention view using the scored triples in `edges_aggregated.csv`. **Note (method).** Interventions were detected by keyword buckets over the non-disease node: vaccination (vaccinat*/inoculat*/lymph), disinfection (disinfect*/phenyl/carbolic/chlor*/fumigat*/limewash), hospital admission (admit*/hospital/ward/in/outpatient/asylum), quarantine/isolation, autopsy/post-mortem, vector control (mosquito/anophe*/larv*/drainage/spray), sanitation (sanitary/latrine/water-supply/clean/scaveng*), and quinine therapy (quinine/cinchona). Diseases follow the matching specification in Section 3.1. We aggregate link strength by the `score` field.

[2]All figures in this paragraph are computed from `relation_topics.csv` by summing the `count` column for the `relation` surface strings (case-insensitive match for the causal family via substring "cause").

Topic	Topic label (illustrative)	Top relations
0	Cultivation / Treatment	All Attempts At Artificial Cultivation Of Have… After After Treatment
1	Plague hospital work	Showed A Low Outturn Of Work Mentioned In The Text At Rs. Do. Do
2	Sanitary departments	TB Medical And Sanitary Departments Of The Government… Treasury On Account Of Establishment Others On
3	Hospital admissions	Lugols Solution Was Tried Resulted From Ages Of Lunatics
4	Autopsies / Inspection	Inspected By Me During The Year Provided For Population Of Were Opened During The Year
5	Sanitation / Analysis	Analysis Of Wool Samples (Bihar) Correlations Amongst Correlated To The Order Of
6	Veterinary services	Conducive To Female Attendants At Rs Indication Of
7	Vaccine preparation	Included In Charges Receives Patients From Imported Opium From
8	Tuberculosis departments	Favour The Multiplication And Development Of Tubercle Bacilli Issues Of Pattern
9	Asylums / Prisons	Central Jail Located In Scale Of Under Observation For

Table 3: Representative topic–relation alignments (top surface-form relations per topic, drawn from `relation_topics.csv`). Longer strings are truncated for readability.

The highest-weight pairs (summed score) highlight familiar interventions: smallpox→vaccination (137.4), plague→hospital admission (63.0), plague→sanitation (46.5), malaria→hospital admission (40.1), and smallpox→hospital admission (25.2).[3] Figure 3 visualizes these top-ranked

[3]Full rankings, including additional pairs such as plague→disinfection and tuberculosis→hospital admis-

Category	Examples (canonicalized)
Diseases	Plague; Cholera; Malaria; Smallpox; Leprosy; Kala-azar
Institutions	Plague hospitals; Tuberculosis departments; Municipal hospitals; Veterinary labs
Interventions	Quinine therapy; Disinfection; Vaccination campaigns; Mosquito control
Relation schemas	Treated_With; Causes; After_Treatment; Vaccinated_Against
Thematic clusters	Epidemics; Hospitals; Vaccination; Sanitation; Veterinary medicine

Table 4: Summary of induced ontology components. The ontology integrates diseases, interventions, institutions, and relation schemas into interpretable clusters.

Theme	Characteristic relations / interventions
Plague	Disinfection protocols; Hospital admission / relapse tracking; Sanitation measures
Malaria	Quinine therapy; Mosquito-vector control; Seasonal pattern reporting
Smallpox	Vaccination campaigns; Dates of inoculation; Geographic coverage
Tuberculosis	Departmental management; Long-term admission; Institutional care
Veterinary medicine	Fowl/pig virus work; Experimental vaccination (animal hosts); Public supply/inspection

Table 5: Induced disease and institutional themes and their characteristic relations/interventions.

disease→intervention pairs by summed `score`. Temporal slices from the same file clarify the shift in priorities.Figure 4 plots these associations across eras (pre-1900, 1900s, 1910s, 1920+), highlighting changes in intervention emphasis over time. [4] Smallpox→vaccination is strong pre-1900 (**85.3**) and persists into the 1900s–1920+ (**19.2**/**16.1**/**15.2**). Plague→hospital admission is concentrated pre-1900 (**28.7**) and then declines (**0.44**/**0.72**/**0.37**); plague→disinfection also peaks pre-1900 (**7.47**). Institutional care rises later: tuberculosis→hospital admission strengthens in the 1910s–1920+ (**4.63**/**2.65**). Vector measures appear around malaria (malaria→vector control **0.38** pre-1900 and **0.92** in 1920+), with quinine therapy surfacing in the 1910s (**0.59**). Taken together, the network recovers a historically plausible sequence: epidemic containment (plague) through admission, sanitation, and disinfection; sustained

sion, are reported in Table 2

[4] Era bins: pre-1900, 1900s, 1910s, 1920+.

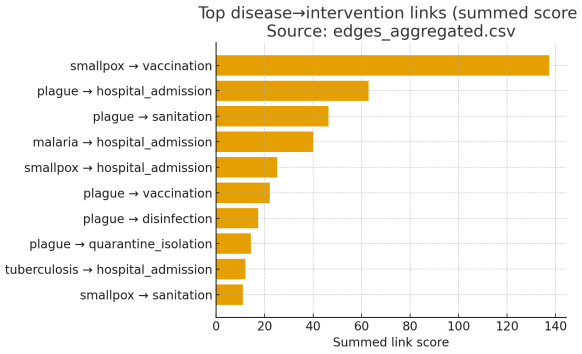

Figure 3: Top-ranked disease→intervention pairs by summed `score`, derived from `edges_aggregated.csv`. Dominant associations include smallpox→vaccination, plague→hospital admission, and malaria→quinine therapy.

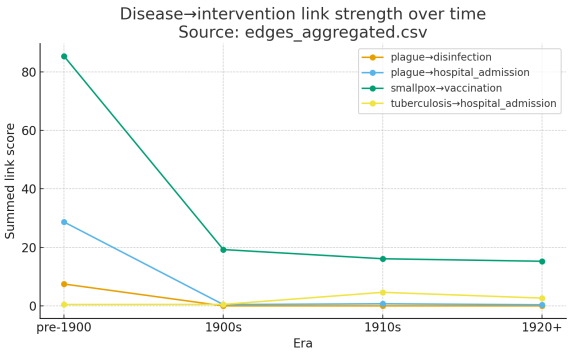

Figure 4: Temporal profiles of selected disease→intervention pairs across four eras (pre-1900, 1900s, 1910s, 1920+), derived from `edges_aggregated.csv`. The trajectories highlight shifts in public health priorities, with plague interventions peaking in the 1900s and tuberculosis care rising after 1910.

vaccination around smallpox; and growing institutional management for tuberculosis, with vector and quinine measures intermittently attached to malaria. The highest-weight pairs (summed score) highlight familiar interventions among the six principal diseases, while tuberculosis is reported separately as an institutional-care trend.

4. Conclusion

We presented a reproducible pipeline for inducing historical biomedical knowledge graphs from noisy OCR text. The approach combines BioBERT NER, LLM-guided triple extraction and validation, canonicalization with schema mapping, BERTopic alignment, and multi-metric edge scoring with temporal decay. Applied to the *Medical History of British India* corpus, the induced resources consolidate a

large surface-form space (22,360 unique relation strings) into interpretable themes (71 clusters), and yield inventories that support both canonicalized schema counts and conservative surface-string tallies.

Analytically, the disease→intervention projection constructed from the scored edges captures historically plausible priorities: strong and persistent *smallpox→vaccination*; crisis-oriented *plague→hospital admission/disinfection* peaking before 1900; and later growth of institutional care signaled by *tuberculosis→hospital admission*. Malaria appears with vector-management and quinine signals where present. These patterns align with topic–relation evidence and case studies, demonstrating that the pipeline recovers both medical practices and the infrastructures (e.g., plague hospitals, tuberculosis departments) through which policy was enacted.

Limitations include residual OCR and administrative noise, imperfect schema normalization for long-tail relations, and the absence of systematic expert historian validation (though resources are released to enable it). Future work will extend to cross-lingual and multilingual historical text collections, refine entity typing and schema coverage, and deepen temporal modeling beyond coarse era bins. These challenges are intrinsic to large-scale historical NLP, but foregrounding them highlights where further interdisciplinary work is needed. While outputs are released for reproducibility, long-term computational reproducibility also depends on stable hosting of large LLMs and biomedical embeddings.

Taken together, the method and released artifacts provide a transparent foundation for large-scale, data-driven study of historical medicine, supporting both NLP reproducibility and future interdisciplinary work (GitHub Repository)

5. Ethics Statement

This study relies exclusively on publicly available historical medical documents from the Medical History of British India corpus (1850–1950). The corpus is digitized and openly released by the National Library of Scotland, and no personally identifiable or sensitive modern data are involved. All analyses were conducted on archival texts, ensuring compliance with ethical standards for research using public-domain resources.

6. References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Seungmin Choi and Yuchul Jung. 2025. Knowledge graph construction: Extraction, learning, and evaluation. *Applied Sciences*, 15(7):3727.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 288–310. Springer.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch. 2012. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Rajat Mishra and S Shridevi. 2024. Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records. *Scientific Reports*, 14(1):25449.

National Library of Scotland. 2019. A medical history of british india. https://data.nls.uk/data/digitised-collections/a-medical-history-of-british-india/.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, 56(11):13071–13102.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662.

Zongbao Yang, Yuchen Lin, Yinxin Xu, Jinlong Hu, and Shoubin Dong. 2023. Interpretable disease prediction via path reasoning over medical knowledge graphs and admission history. *Knowledge-Based Systems*, 281:111082.