



Standardization and Harmonization of EHR data for clinical phenotyping and patient graph construction with OMOP Common Data Model and Biomedical Ontologies

Bableen Kaur^{1,2} **Sagarika**^{2,3} **Gautam Ahuja**^{1,2,4} **Siddhant Poudyal**^{1,2} **Himani Balutia**^{1,2,5} **Sonali Bhat**^{1,2,5} **Gracy Singh**^{1,2,5} **Vinita Sharma**^{2,3} **Ashutosh Yadav**⁶ **Rajesh Saxena**⁶ **Swati Waghdhare**⁶ **Vinitaa Jha**^{6,7} **Anurag Agrawal**^{1,3,8} **Rintu Kutum**^{1,2,4,8,9}

¹Koita Centre for Digital Health at Ashoka, Ashoka University ²Augmented Health Systems Laboratory, Ashoka University ³Department of Biology, Ashoka University ⁴Mphasis AI & Applied Tech Lab at Ashoka, Ashoka University ⁵Banasthali Vidyapith ⁶Max Healthcare ⁷Max Healthcare Research & Academics ⁸Trivedi School of Biosciences, Ashoka University ⁹Department of Computer Science, Ashoka University

Background

This poster outlines a systematic workflow for extracting, transforming and loading unstructured clinical data (EHR) specific to the Indian context into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) through modeling data schema, data preprocessing, standardization and harmonization of clinical terms. This systematic workflow demonstrates the development of a patient-centric knowledge graph using contextual CDM tables, enabling data integration, semantic querying and advanced machine learning model implementation, such as discovering patient subgroups from standardised and harmonized EHR data.

Methods and Results

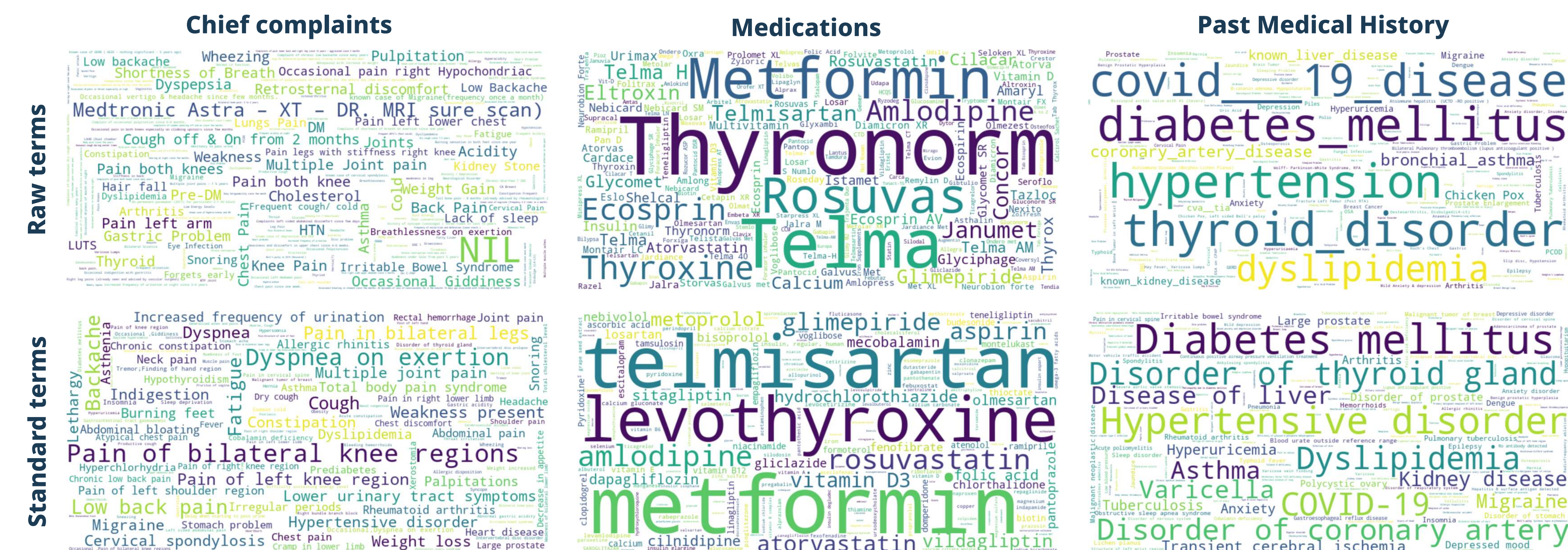


Figure: Word clouds comparing raw EHR terminology (top row) with mapped OMOP CDM standard concepts (bottom row) across chief complaints (from 872 to 582 terms), medications (from 3,685 to 525 terms), and past medical history (from 5,252 to 176 terms). Font size reflects term frequency in the EHR, demonstrating consolidation and reduced variability after standardization.

Category-wise Allergen Mapping

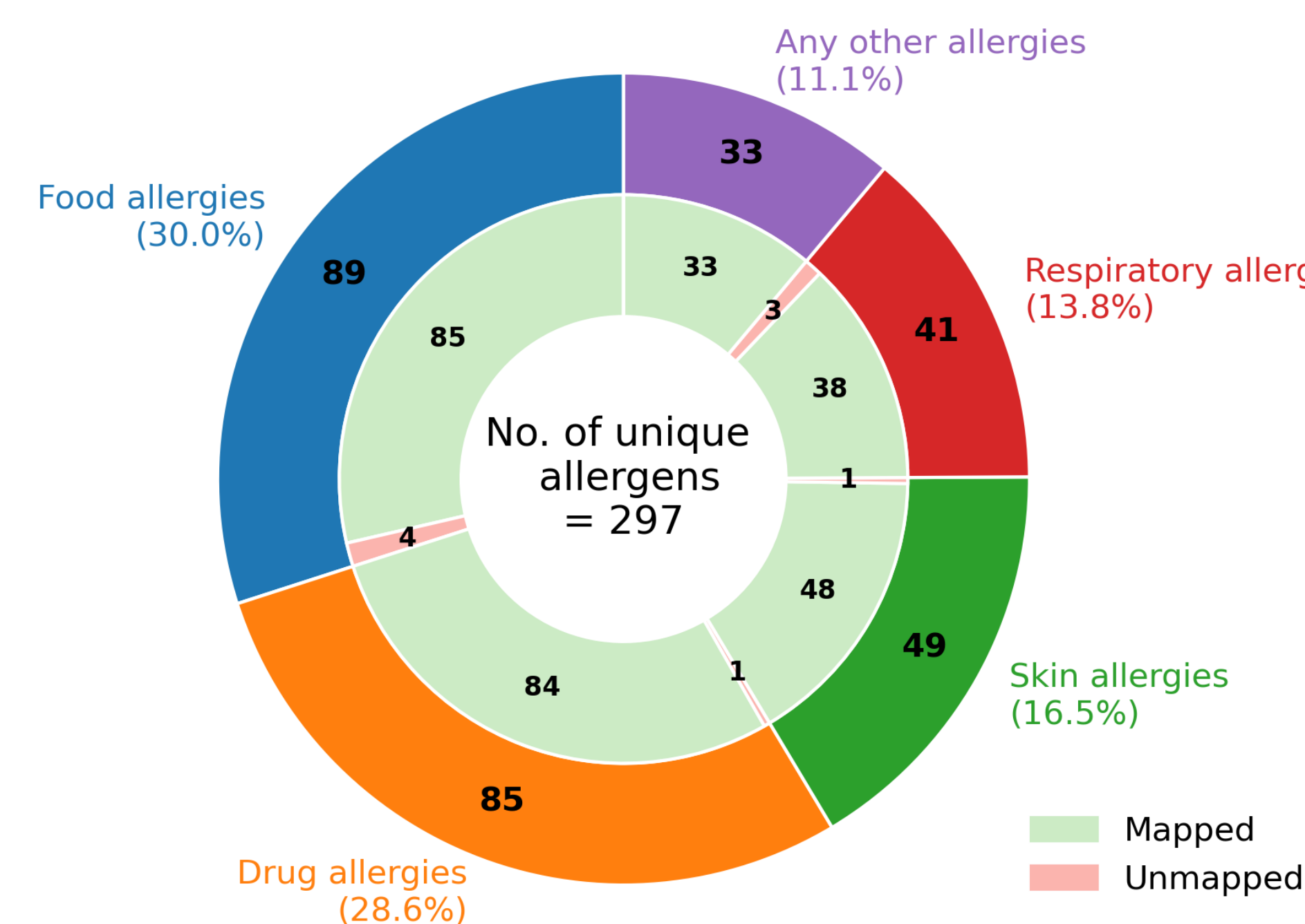


Figure: **Category-wise distribution of unique allergens recorded in the EHR and their mapping status.** The outer ring shows the total number of unique allergens reported per category, while the inner ring depicts whether these allergens were successfully mapped to OHDSI standardized vocabularies or remained unmapped. In total, 297 unique allergens across five categories were analyzed.

Figure: **Mapped allergen summary by mapping approach and standardization.** Allergens were mapped either directly to standardized concepts or through manual annotation. The sankey plot shows how mapped allergens from each category progress through the mapping methods and ultimately to either OMOP CDM standardized or non-standard concepts. In total, 288 allergens were successfully mapped.

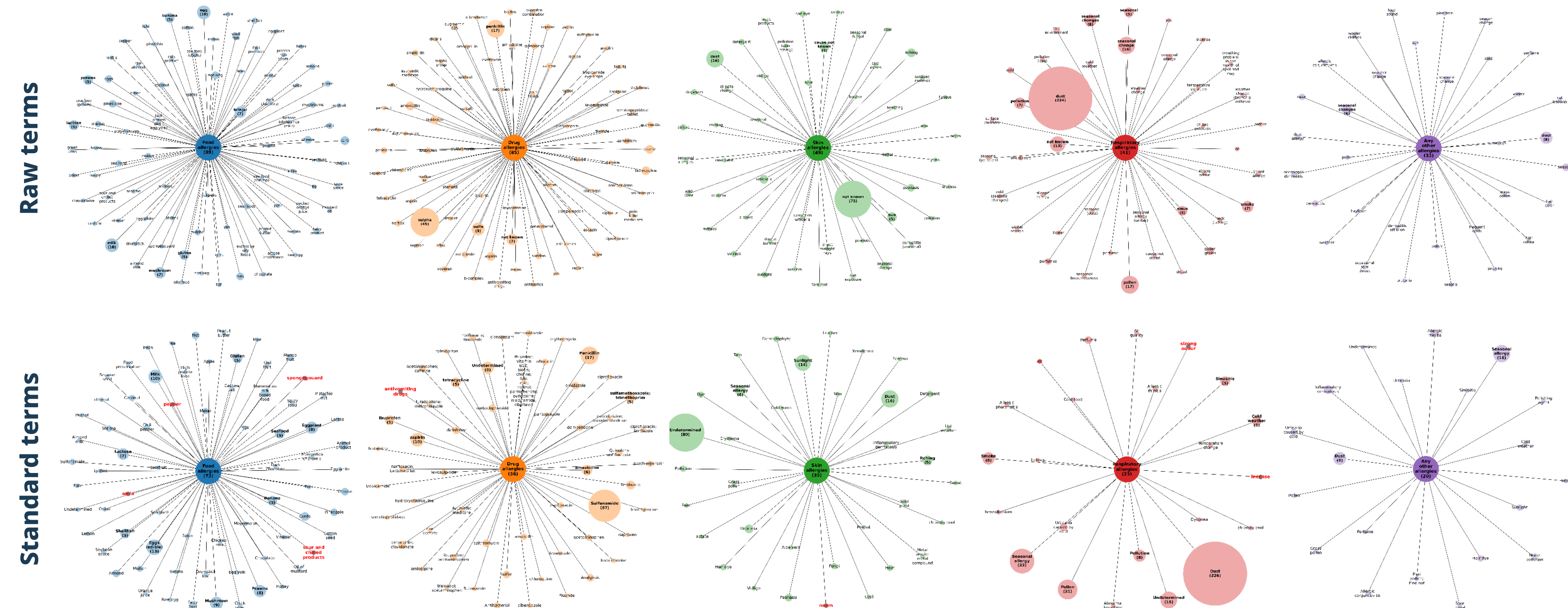
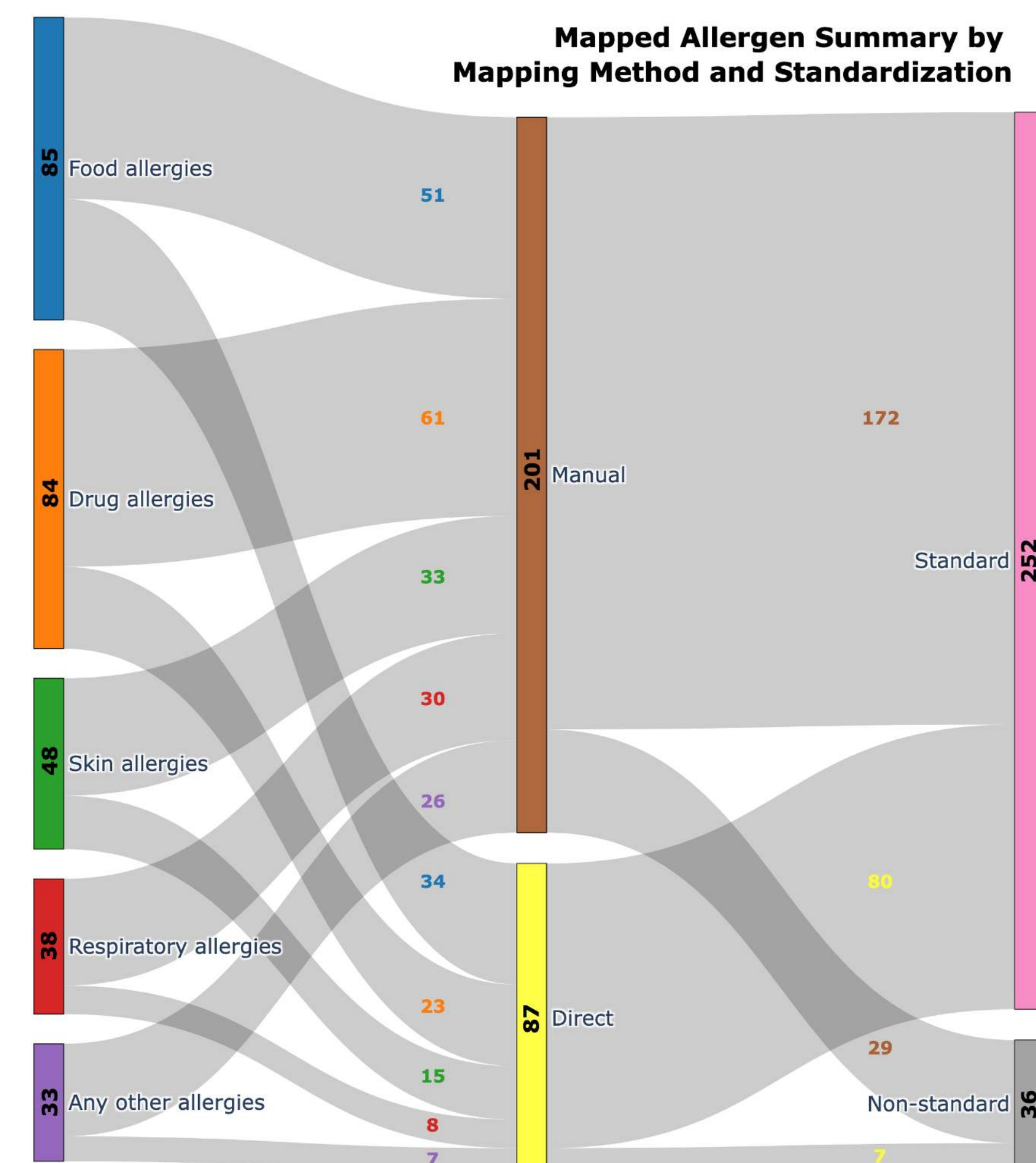


Figure: **Category-wise visualization of allergen terminology in the EHR before and after standardization.** In each category, the graph on the top shows raw allergen terms extracted from the EHR and graph on the bottom shows those successfully mapped to OMOP CDM standard concepts. These paired visualizations illustrate the variability of clinical terminology and the improvements achieved through standardized harmonization.

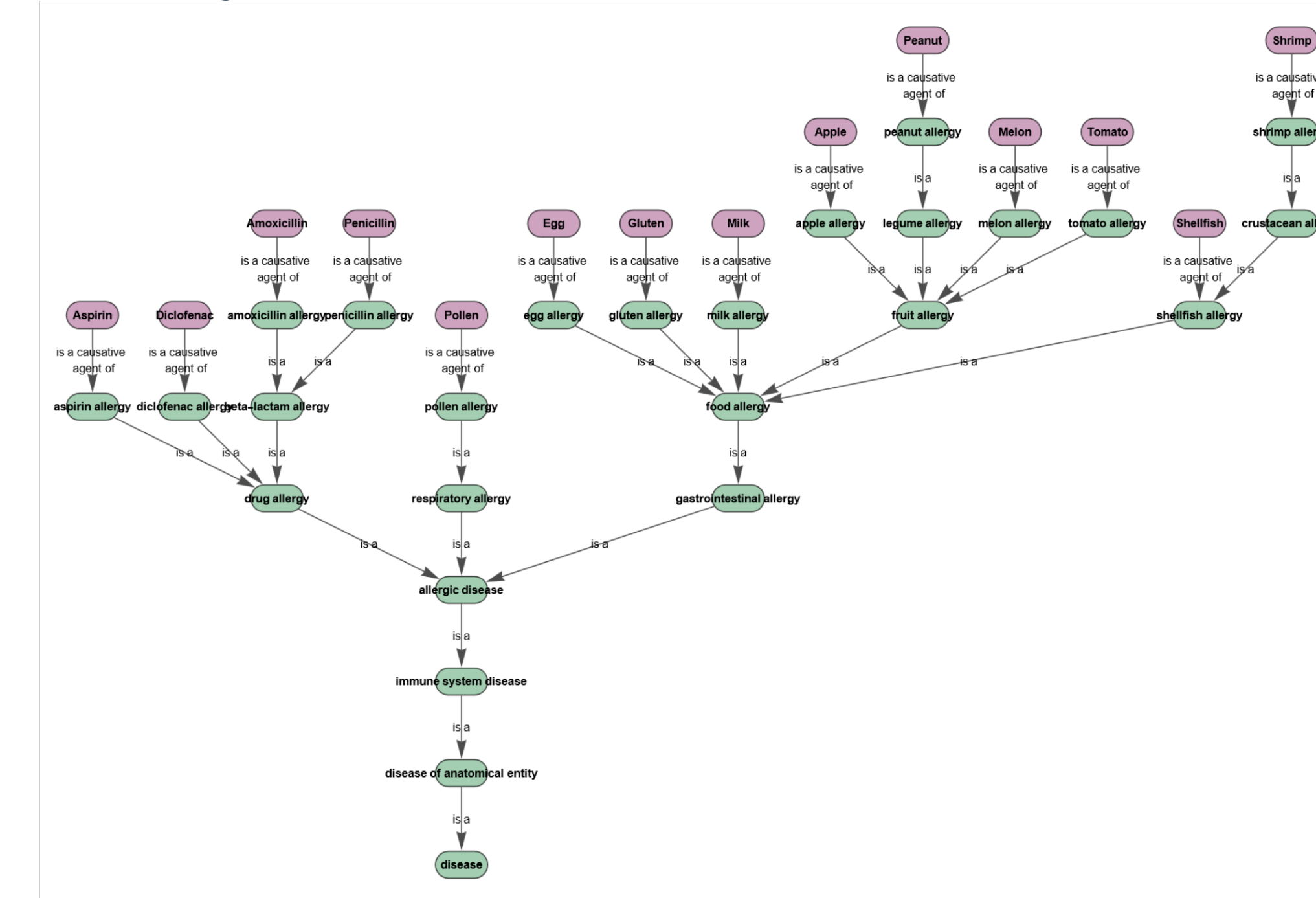


Figure: **Direct mapping of raw EHR-recorded allergen terms to Disease Ontology (DO).**

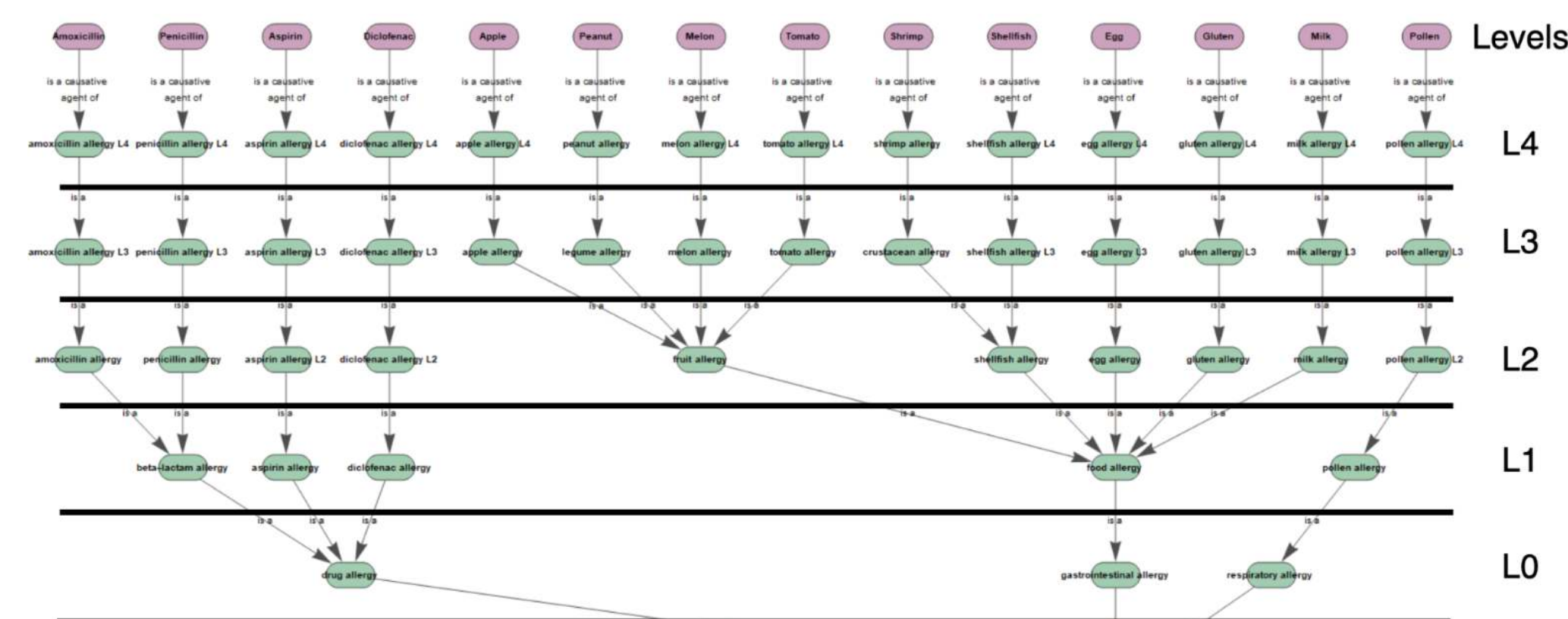


Figure: **Hierarchy of standardized allergen concepts mapped to Disease Ontology (DO),** organized by level (L0-L4) based on structural distance from the root disease class. This level-wise representation demonstrates the positioning of mapped standardized concepts within the broader disease taxonomy and emphasizes the improved contextualization achieved through standardization.

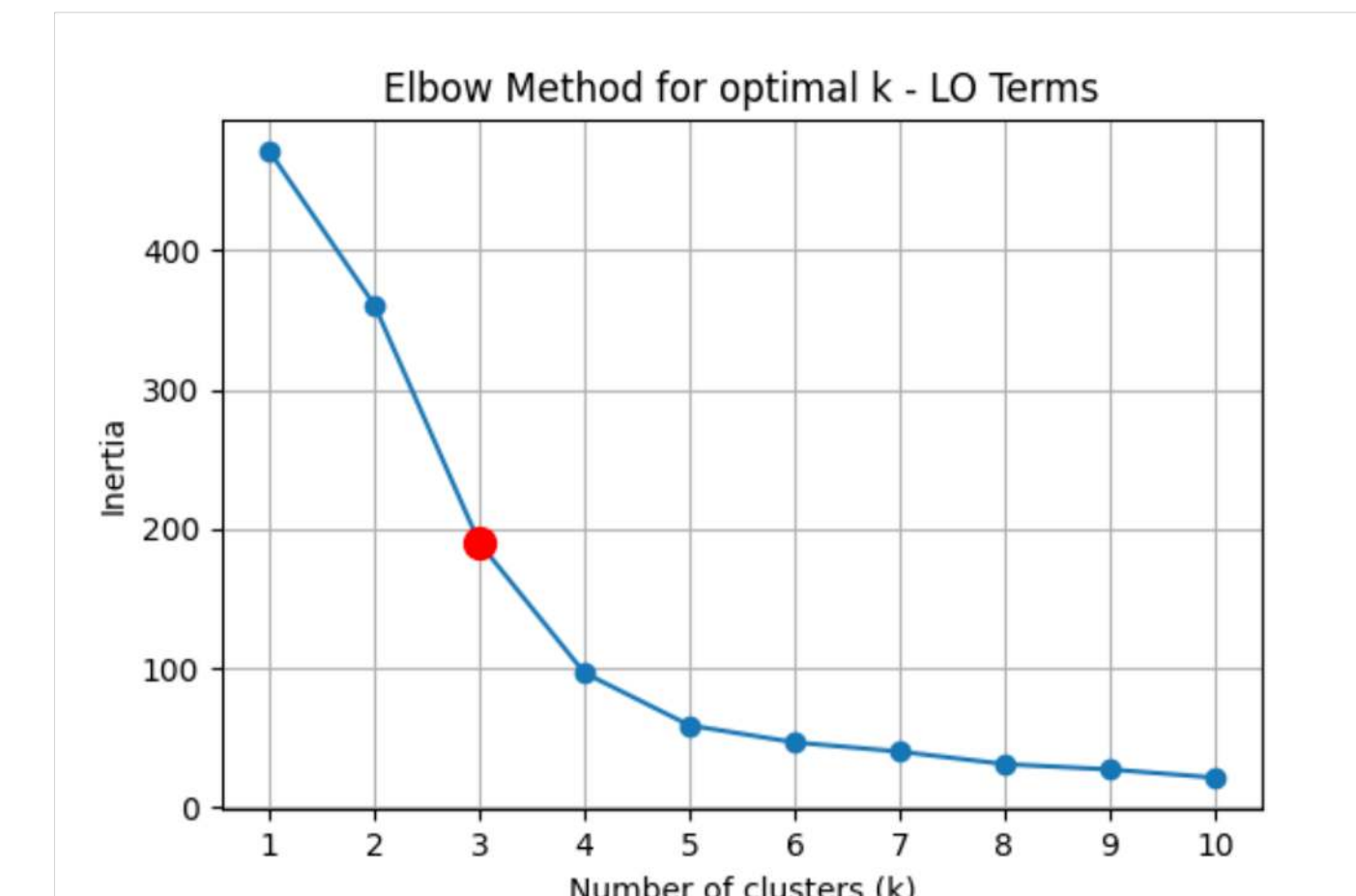
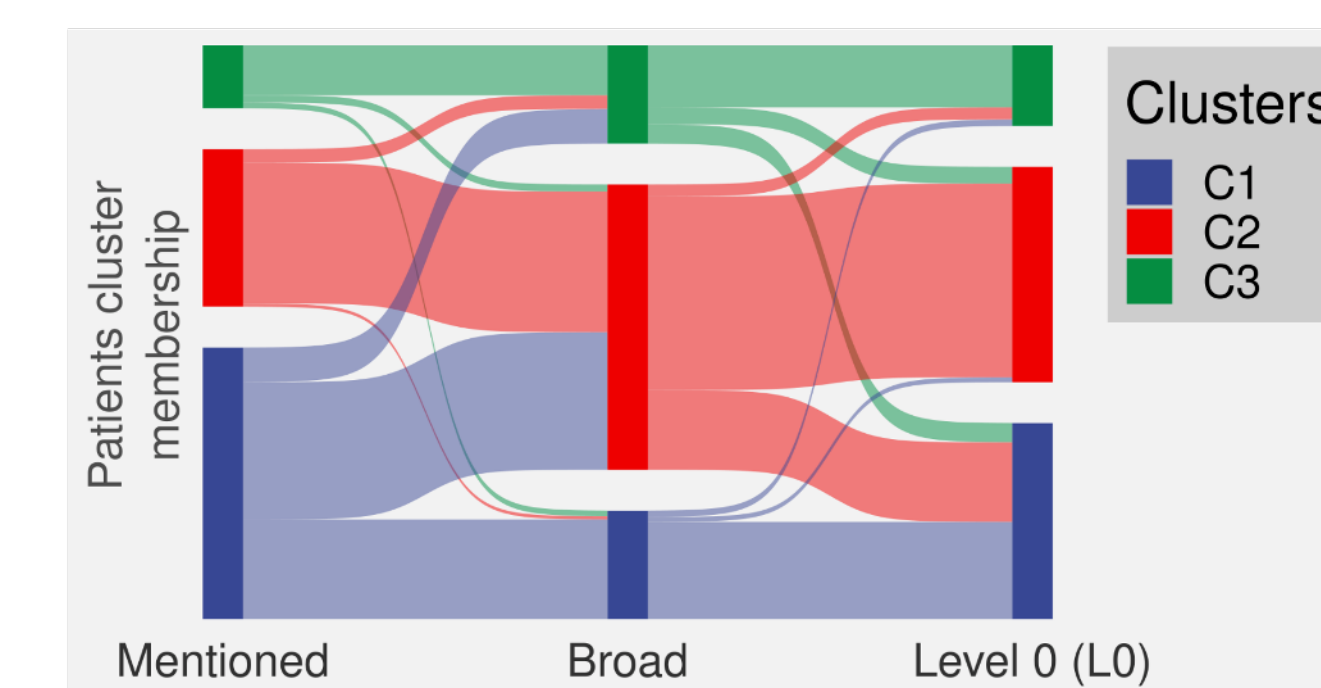
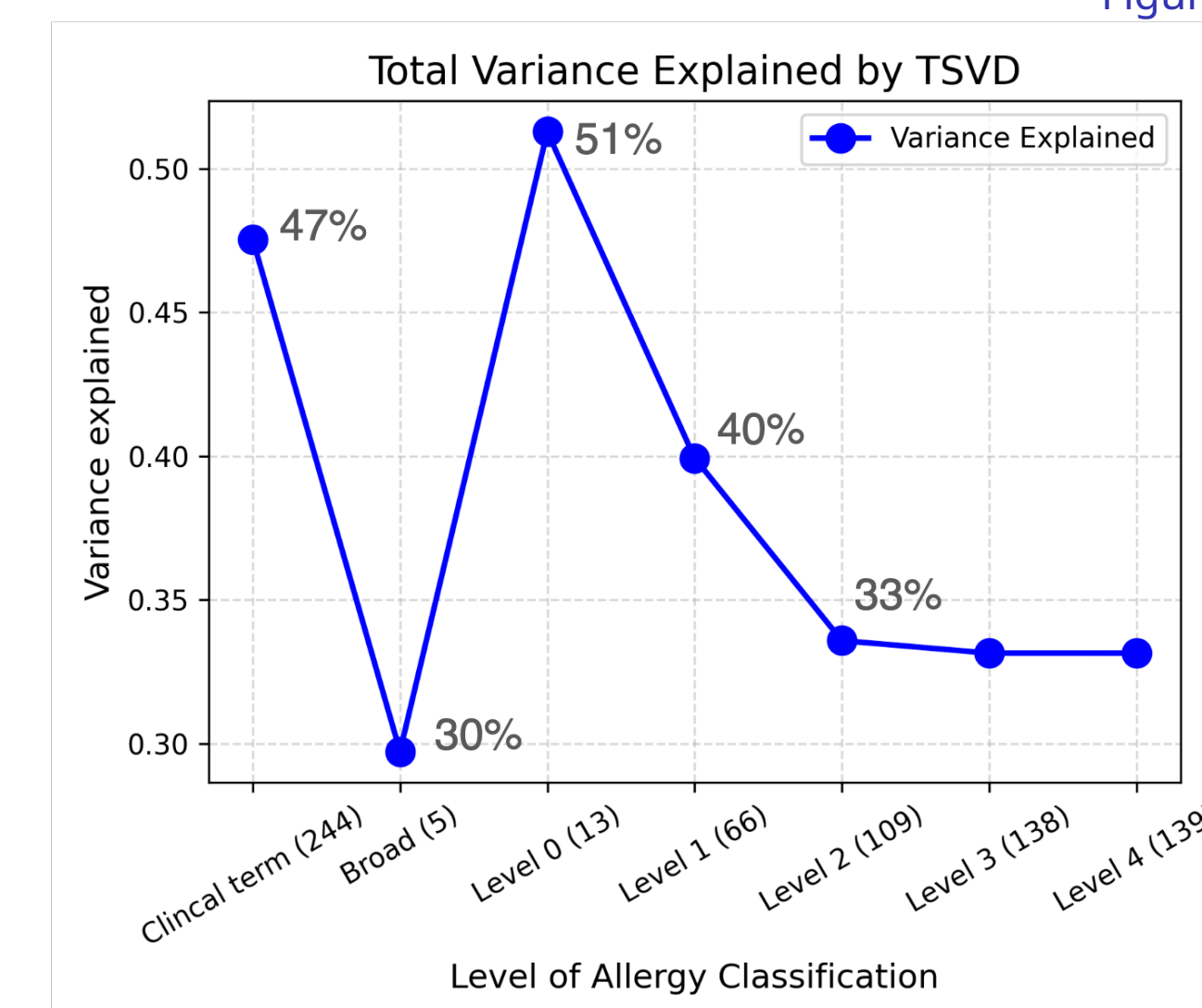


Figure: **Dimensionality reduction and clustering assessment across terminology abstraction levels.** Truncated Singular Value Decomposition (TSVD) was applied to the constructed incidence matrices to estimate the variance explained at each level of allergy term representation, with the highest retention observed at Level 0 (51%) (left). The Sankey diagram illustrates transitions in patient cluster membership when shifting from raw mentioned terms to Broad category terms and then to Level 0 standardized concepts (middle). The Elbow method identified $k = 3$ as the optimal number of clusters for Level 0 terms based on inertia reduction (right).

Conclusions

To estimate the variance explained based on the incidence matrix, we performed truncated SVD due to the sparseness of the matrix. The higher percentage of variance explained indicates rich information content in the matrix, indirectly indicating an optimal hypergraph for unstructured EHR data. For Level 0 - L0 (13 hyperedges), we find cumulative variance explained (top 2) of 51.29% in comparison to the mentioned clinical terms (29.72%) and for broad terms (47.5%). This highlights that hypergraphs over interoperable clinical terms have higher information content than those over unstructured clinical terms. We would like to model hypergraph with more clinical data.