Original Paper

# Using Structured Codes and Free-Text Notes to Measure Information Complementarity in Electronic Health Records: Feasibility and Validation Study

Tom M Seinen, MSc; Jan A Kors, PhD; Erik M van Mulligen, PhD; Peter R Rijnbeek, PhD

Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands

**Corresponding Author:**
Tom M Seinen, MSc
Department of Medical Informatics
Erasmus University Medical Center
Dr. Molewaterplein 40
Rotterdam, 3015 GD
Netherlands
Phone: 31 010 7044122
Email: t.seinen@erasmusmc.nl

## Abstract

**Background:** Electronic health records (EHRs) consist of both structured data (eg, diagnostic codes) and unstructured data (eg, clinical notes). It is commonly believed that unstructured clinical narratives provide more comprehensive information. However, this assumption lacks large-scale validation and direct validation methods.

**Objective:** This study aims to quantitatively compare the information in structured and unstructured EHR data and directly validate whether unstructured data offers more extensive information across a patient population.

**Methods:** We analyzed both structured and unstructured data from patient records and visits in a large Dutch primary care EHR database between January 2021 and January 2024. Clinical concepts were identified from free-text notes using an extraction framework tailored for Dutch and compared with concepts from structured data. Concept embeddings were generated to measure semantic similarity between structured and extracted concepts through cosine similarity. A similarity threshold was systematically determined via annotated matches and minimized weighted Gini impurity. We then quantified the concept overlap between structured and unstructured data across various concept domains and patient populations.

**Results:** In a population of 1.8 million patients, only 13% of extracted concepts from patient records and 7% from individual visits had similar structured counterparts. Conversely, 42% of structured concepts in records and 25% in visits had similar matches in unstructured data. Condition concepts had the highest overlap, followed by measurements and drug concepts. Subpopulation visits, such as those with chronic conditions or psychological disorders, showed different proportions of data overlap, indicating varied reliance on structured versus unstructured data across clinical contexts.

**Conclusions:** Our study demonstrates the feasibility of quantifying the information difference between structured and unstructured data, showing that the unstructured data provides important additional information in the studied database and populations. The annotated concept matches are made publicly available for the clinical natural language processing community. Despite some limitations, our proposed methodology proves versatile, and its application can lead to more robust and insightful observational clinical research.

***(J Med Internet Res 2025;27:e66910)*** doi: 10.2196/66910

**KEYWORDS**

natural language processing; named entity recognition; clinical concept extraction; machine learning; electronic health records; EHR; word embeddings; clinical concept similarity; text mining; code; free-text; information; electronic record; data; patient records; framework; structured data; unstructured data

## Introduction

Electronic health records (EHRs), originally designed for clinical documentation and administration, are now increasingly used in observational research [1,2], supporting various types of studies, including case studies, patient cohort characterizations, and clinical prediction modeling. EHR data are generally recorded in 2 forms: structured and unstructured data. Structured data includes clinical codes for documenting clinical events, such as diagnoses, medications, procedures, and measurements. Structured data is particularly suitable for observational research due to its consistent meaning, tabular format, and standardized vocabulary of codes. Unstructured data consists of free-text clinical notes, which can provide detailed descriptions capturing the nuances of patient care, such as physician observations, patient histories, diagnostic impressions, and discharge summaries. Although rich in contextual information, unstructured data poses challenges for direct analysis because of its variability and lack of standardization [3,4]. Consequently, extracting meaningful information from unstructured data requires significant investment in manual labor, computational resources, and time.

It is commonly assumed that the text data contains more detailed and extensive information than structured data, based on the often-reported claim—grounded in business-related data [5]—that 80% of EHR data is unstructured [3,4,6-8]. While this assumption may hold, its validity is influenced by factors like documentation quality and clinical context. For example, intensive care data is characterized by a high frequency of measurements, while psychiatric care relies more on textual narratives. Many studies explored the added value of information from text by comparing analyses with and without it [9-15], indirectly validating this assumption. However, even if the assumption holds, it initially remains uncertain to what extent the information in the text data matches or complements the structured data.

Understanding the quantity and differences in the information available in structured and unstructured data for a specific database offers several advantages for observational clinical research. First, it aids study design by identifying the most abundant and reliable data types, enabling researchers to formulate feasible hypotheses and research questions. Second, it allows for more effective allocation of human and computational resources by focusing efforts where they are most needed. Third, knowing the balance between structured and unstructured data helps researchers prioritize according to the study's specific needs. Finally, it highlights gaps or unique aspects of the data, facilitating the exploration of underused research opportunities.

Comparing the information from structured and unstructured data involves various measures, such as quantity and content. While structured data points can be counted and unstructured data quantified by individual words or extracted concepts, comparing content similarity is more challenging. The core meaning of both structured codes and unstructured text lies in their semantic content. Evaluating the information distance between 2 concepts or texts requires comparing their semantic meanings, a task commonly addressed in natural language processing through semantic similarity measures. Modern approaches often use word embedding models to generate concept embeddings, which are used in applications like biomedical ontology matching [16,17] and concept normalization [18-20]. Specialized embedding models, such as SapBERT [21] and BioLORD [22], have been developed for this purpose and provide the opportunity to measure the information difference between structured and unstructured data.

Several studies have compared structured and unstructured data for specific clinical variables, such as social and behavioral determinants of health [23,24] and smoking history [25,26]. However, to our knowledge, no research has directly assessed the information differences between structured and unstructured data across all clinical events in a database. Our study aims to quantitatively compare the information coded by general practitioners (GPs) with the information documented in free-text notes, using data from a large Dutch GP database. We extracted clinical concepts from unstructured text and used concept embeddings to calculate their similarity with the structured concepts. After determining a similarity threshold, we estimated the difference and overlap of information between structured and unstructured data.

## Methods

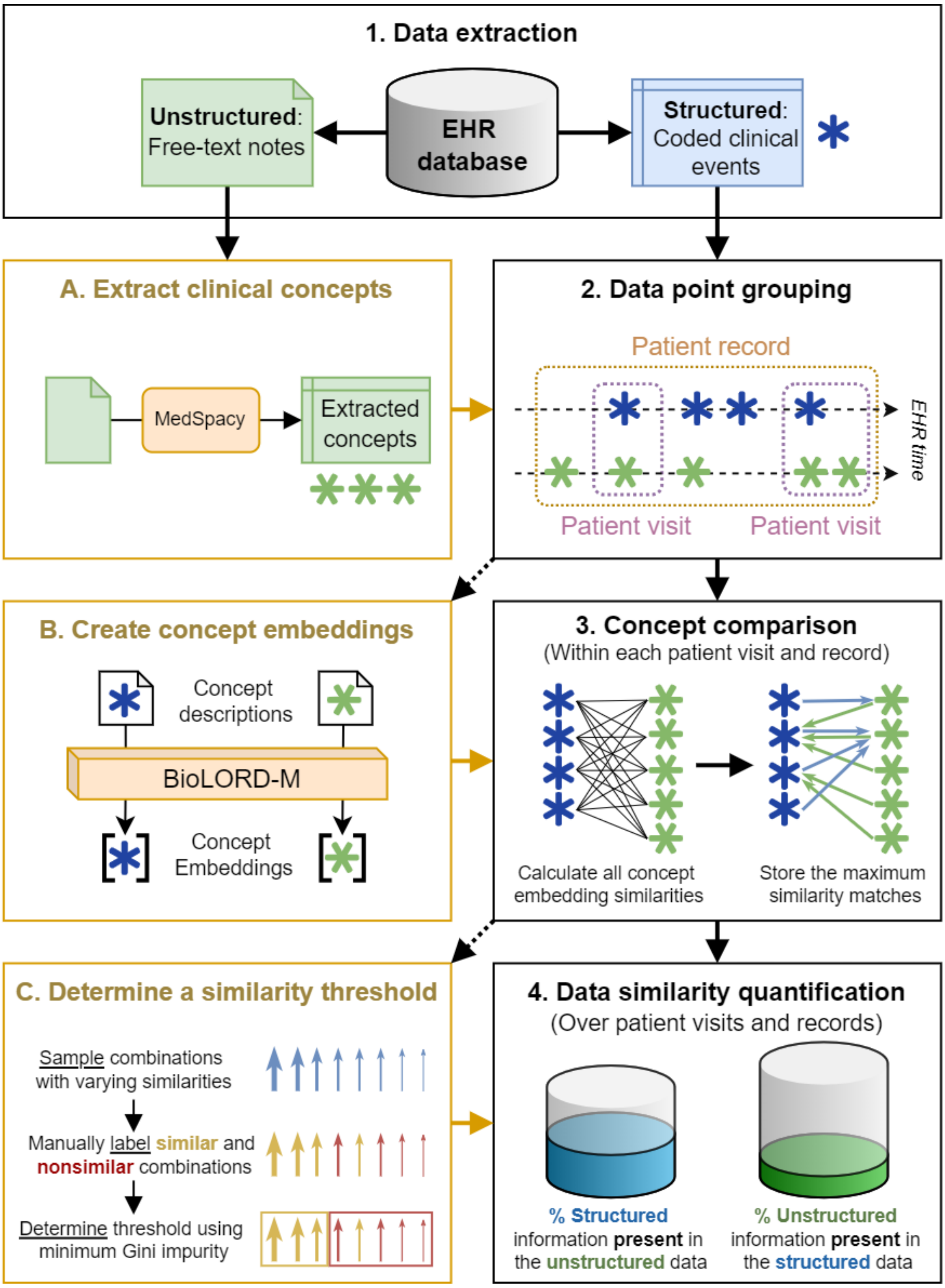
### Database and Setting

This study used the Integrated Primary Care Information (IPCI) database [27], a longitudinal EHR database of Dutch GPs. IPCI contains records of 2.9 million patients with a median follow-up of 4.8 years, spanning from 1993 to 2024. The database is standardized using the Observational Medical Outcomes Partnership Common Data Model [28]. Eligible participants in our study dataset included all patients recorded in the database from January 2021 to January 2024. The study received approval from the IPCI governance board under code 2023-04.

### Methodological Setup

The methods consist of 4 main parts, visualized in Figure 1, each described in detail in the following sections. First, we extracted both structured and unstructured data for each eligible patient and applied a concept extraction framework to extract clinical concepts from the free-text notes. Second, we applied 2 different data grouping methods to the population. Third, we used pretrained multilingual concept embeddings to calculate the similarities between the structured and extracted concepts. Finally, we annotated a sample of concept similarity matches to determine a similarity score threshold and quantified the data similarity in the database.

**Figure 1.** Visualization of the methodological setup. Steps 1-4 outline the primary process, including data extraction, data point grouping, concept comparison, and data similarity quantification. Steps A, B, and C detail additional processes, specifically the extraction of clinical concepts, the creation of concept embeddings, and the determination of a similarity threshold. EHR: electronic health record.


1. Data extraction
Unstructured: Free-text notes
EHR database
Structured: Coded clinical events
A. Extract clinical concepts
MedSpacy
Extracted concepts
2. Data point grouping
Patient record
EHR time
Patient visit
Patient visit
B. Create concept embeddings
Concept descriptions
BioLORD-M
Concept Embeddings
3. Concept comparison
(Within each patient visit and record)
Calculate all concept embedding similarities
Store the maximum similarity matches
C. Determine a similarity threshold
Sample combinations with varying similarities
Manually label similar and nonsimilar combinations
Determine threshold using minimum Gini impurity
4. Data similarity quantification
(Over patient visits and records)
% Structured information present in the unstructured data
% Unstructured information present in the structured data

## Data Extraction

We extracted all structured and unstructured data points for each eligible patient (step 1 in Figure 1). Structured data includes conditions, procedures, prescriptions, measurements, and observations. The coding systems used, such as the International Classification of Primary Care-1, are listed in Table S1 in Multimedia Appendix 1. Unstructured data consists of 3 types of notes: subjective, objective, assessment, and plan notes from consultations; referral and communication notes with secondary care; and other notes from free-text fields in the EHR system, primarily accompanying condition codes.

## Clinical Concept Extraction

Structured coded clinical events are considered single data points, while free-text notes can contain multiple pieces of information embedded in the narrative. To compare these, we extracted individual data points from unstructured text using clinical named entity recognition and linking, generally known as clinical concept extraction (step A in Figure 1). We used MedSpacy [29], a toolkit that extracts clinical concepts based on a reference thesaurus such as the unified medical language system (UMLS) and detects contextual modifiers using language-specific rules. We used a version of MedSpacy adapted for Dutch [10,30], incorporating all Dutch vocabularies from UMLS and replacing the English Systematized Nomenclature of Medicine Clinical Terms with the Dutch translation [31]. We used Dutch context rules to detect qualifiers such as negation, temporality, and experiencer. These rules were previously validated with an annotated corpus [32]. The extracted UMLS concepts were mapped to concept domains such as observation, condition, measurement, drug, and procedure in the Observational Medical Outcomes Partnership standardized vocabulary, which contains the majority of UMLS vocabularies, for easy comparison with the structured data.

## Data Point Grouping

Given the longitudinal nature of GP EHR data, we used 2 grouping methods for comparison: by patient visit and by patient record (step 2 in Figure 1). Grouping by visit considers data points recorded simultaneously during a patient visit, providing a natural basis for comparison, while grouping by record includes additional information recorded outside visits, such as lab results and secondary care communications, allowing for a broader comparison.

## Concept Comparison

Each structured or extracted data point is represented by a single clinical concept. We compared the Cartesian product of structured and extracted concepts within each group (step 3 in Figure 1). For example, a visit with 1 coded condition and 1 prescription, along with a subjective, objective, assessment, and plan note containing 5 extracted concepts, results in 10 comparisons. While we could count exact concept matches, 2 issues arise. First, the compared concepts can be very similar but not identical, for example, the International Classification of Primary Care-1 code “Fracture: hand/foot bone” (L74) and the Systematized Nomenclature of Medicine Clinical Terms concept “Closed fracture of hand” (704005005). Second, the concept vocabularies differ, as the GPs choose from a limited set of codes, while concept extraction uses the complete UMLS.

To overcome these issues, we used concept embeddings for fuzzy matching, enabling us to measure semantic similarities and recognize similar concepts. Cosine similarity, which measures the angle between 2 embedding vectors, was used to calculate similarities, ranging from –1 (strongly opposite) to 1 (very similar). For each structured and extracted concept, we stored only the highest similarity match, as we are only interested in finding the most similar concept in the other data type.

## Concept Embeddings

We used the BioLORD-2023-M pretrained sentence transformer model [22] to generate the multilingual concept embeddings, as visualized in step B of Figure 1. BioLORD-2023-M is designed to produce meaningful representations for biomedical concepts across multiple languages, including English and Dutch, and its cross-lingual performance has been evaluated by the authors [22]. By inputting the concept description, from either UMLS for the extracted concepts or from the source vocabulary for the structured concepts, the model generates a dense 768-dimensional vector, allowing us to create embeddings for both structured and extracted concepts. For concepts with multiple descriptions or synonyms, we calculated an embedding for each and averaged them to create one comprehensive concept representation. For structured observations and measurements, we included the unit and value in the description to enrich the embedding. The model's multilingual capability enabled us to embed concept descriptions in both Dutch and English within the same latent space.

## Similarity Threshold Determination

To quantify the information difference between structured and extracted concepts, we needed to define a threshold for concept similarity. Since concept similarity depends on the nature of the embeddings, we developed a systematic method to determine this threshold, as visualized in step C of Figure 1. First, we randomly selected concept pairs at various similarity levels, ranging from a similarity score of 0.35 to 1, with samples taken at 0.05 intervals. This sampling was done for both structured and extracted concepts across patient visits and records, ensuring each concept domain was represented by sampling 5 concepts per domain. Next, we manually annotated the concept pairs as either similar or nonsimilar. Using these annotations, we determined the threshold at which the weighted Gini impurity [33] of the split between similar and nonsimilar matches was lowest.

## Data Similarity Quantification

Using a similarity threshold and the most similar counterpart for each structured and extracted concept, we determined the number of structured concepts found in free-text (structured-to-unstructured) and the number of extracted concepts that were coded in the structured data (unstructured-to-structured), as shown in step 4 of Figure 1. It is important to note that these counts are not reciprocal since we consider the maximum similarity per concept. For example, multiple extracted concepts from a patient visit text may be highly similar to a single structured concept, but we only compare the structured concept to its most similar extracted counterpart to determine its presence in the text, not the frequency. We calculated these counts and their percentages across the entire set of concepts, as well as within different concept domains, to explore domain-specific differences. For extracted concepts, we only included those without context modifiers to focus on the core unmodified concepts, which ensures a higher degree of certainty and a more straightforward, reliable comparison.

## Subpopulation Comparison

While observing data similarity across the entire population or all GP visits is insightful, applying this method to smaller subpopulations may provide further detail. We defined 3 subpopulations based on different types of clinical events: visits for chronic disorder (type 2 diabetes mellitus), acute event (COVID-19 vaccination), and psychological disorder (depression), as detailed in Table S2 in Multimedia Appendix 1. We then quantified the similarity between structured and unstructured data for these subpopulations, similar to the full population.

# Results

## Population and Data Characteristics

Table 1 presents the population and data characteristics for patient records and visits, including details per observation. Additionally, characteristics are listed for 3 subpopulations. Key differences between patient visits and records are evident. First, each patient has one complete record but can have multiple visits. Second, the total number of structured events and their median values indicate that many events are recorded outside GP visits, as their total numbers are roughly 3 times higher in records. Similarly, the total number of clinical notes is twice as high in records compared to visits alone. Third, more extracted concepts per structured event are generally found in patient visits, but the median number of concepts per note is the same.

**Table 1.** Population and data characteristics for patient records, patient visits, and 3 subpopulation-specific visits.

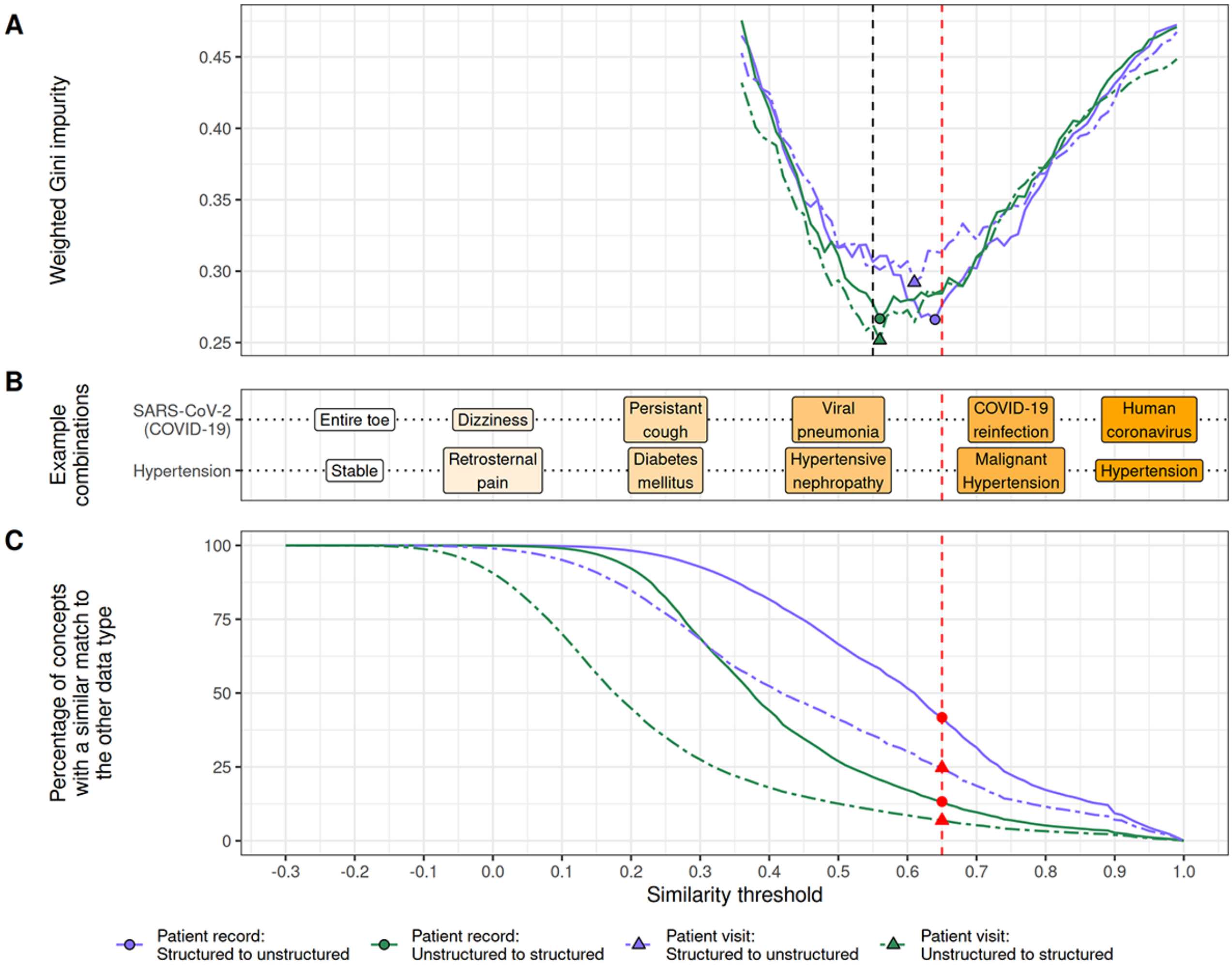
| Characteristic | Patient records | Patient visits | COVID-19 vaccination visits | Depression visits | Diabetes visits |
|---|---|---|---|---|---|
| **Total numbers** | | | | | |
| Number of patients | 1,794,209 | 1,507,473 | 62,724 | 20,519 | 51,032 |
| Number of observations | 1,794,209 | 13,995,524 | 84,484 | 63,957 | 324,698 |
| Ratio number of observations/number of patients | 1 | 9.28 | 1.35 | 3.12 | 6.36 |
| Number of structured events | 132,287,566 | 41,232,407 | 281,606 | 118,351 | 4,513,756 |
| Number of free-text notes | 95,790,377 | 47,915,131 | 188,318 | 216,449 | 1,253,418 |
| Number of extracted concepts | 635,232,943 | 245,052,660 | 688,420 | 1,374,430 | 5,329,978 |
| Ratio extracted concepts/structured events | 4.80 | 5.94 | 2.44 | 11.61 | 1.18 |
| **Statistics per observation** | | | | | |
| Percentage male, % | 49 | 49 | 45 | 33 | 51 |
| Median age (years) | 39 | 39 | 62 | 49 | 68 |
| Median number of structured events | 33 | 2 | 2 | 1 | 6 |
| Median number of free-text notes | 35 | 3 | 2 | 3 | 3 |
| Median number of extracted concepts | 172 | 12 | 5 | 15 | 12 |
| Median number of concepts per note | 3 | 3 | 2 | 3 | 3 |

Characteristics for visits across different clinical events also show clear contrasts. As expected, visits regarding diabetes presented the most visits per patient, while for COVID-19 vaccination, the number of visits per patient was the least. Furthermore, diabetes visits contain much structured and unstructured data, whereas depression visits rely mostly on unstructured data with a large median number of extracted concepts per visit and few structured events. Demographic differences are also notable: diabetes and vaccination populations are similar, but the depression population consists of younger females.

## Determining the Similarity Threshold

We annotated 1764 matches in 4 concept samples: structured-to-unstructured and unstructured-to-structured in both patient record and visit populations. For transparency and reproducibility, the annotated samples are available in Table S3 in Multimedia Appendix 2. Figure 2B visualizes some example extracted concepts matched with increasing similarity to the structured concepts SARS-CoV-2 (COVID-19) and hypertension. The weighted Gini impurity calculated over the binary annotated matches at different similarity thresholds is presented in Figure 2A. We found the minimum impurity of each of the 4 samples between thresholds of 0.55 and 0.65. To be conservative, we set the threshold at 0.65 for data similarity quantification.

**Figure 2.** Analysis across similarity thresholds. (A) The weighted Gini impurity measures the level of impurity in the annotated concept combinations in the 4 samples over the different thresholds. The points indicate the minimum impurity in each sample, and the dashed lines represent their lower bound (0.55) and upper bound (0.65). (B) Examples of extracted concepts matched to the structured concepts of SARS-CoV-2 (COVID-19) and hypertension across different similarity thresholds. (C) The percentage of structured concepts matched the extracted concepts (blue) over the range of similarity thresholds and vice versa (green) for both patient records and visits. The red dashed line in all figures depicts the determined similarity threshold value of 0.65.



## Similarity of Structured and Unstructured Data

The percentage of structured concepts that have a similar match to an extracted concept and vice versa over various similarity thresholds is visualized in Figure 2C for both patient visits and records. At the similarity threshold of 0.65, indicated by the red line in Figure 2, 42% (55.1 million of 132.1 million) of structured concepts in patient records are similar to a concept extracted from text, compared to 25% (9.3 million of 37.8 million) for visits. In contrast, only 13% (66.9 million of 501.9 million) of extracted concepts are similar to a structured concept in patient records, and 7% (11.3 million of 155.7 million) in visits (Figure 3A). This indicates that information in the structured data is more likely to be present in the unstructured data than vice versa. The difference between patient records and visits can be explained by the number of available concepts for matching the entire record versus a single visit.

**Figure 3.** Comparison of structured and extracted concepts in patient records and visits. (A) Total number of structured (left) and extracted concepts (right), along with the number of concepts matched at a similarity threshold of 0.65 (blue for structured to unstructured data, green for unstructured to structured). The percentage of matched concepts is listed in the chart for each data type. (B) The proportion of each concept domain contributed to the total number of structured (left) and extracted concepts (right), along with the proportion of concepts in each domain matched to the other data type.

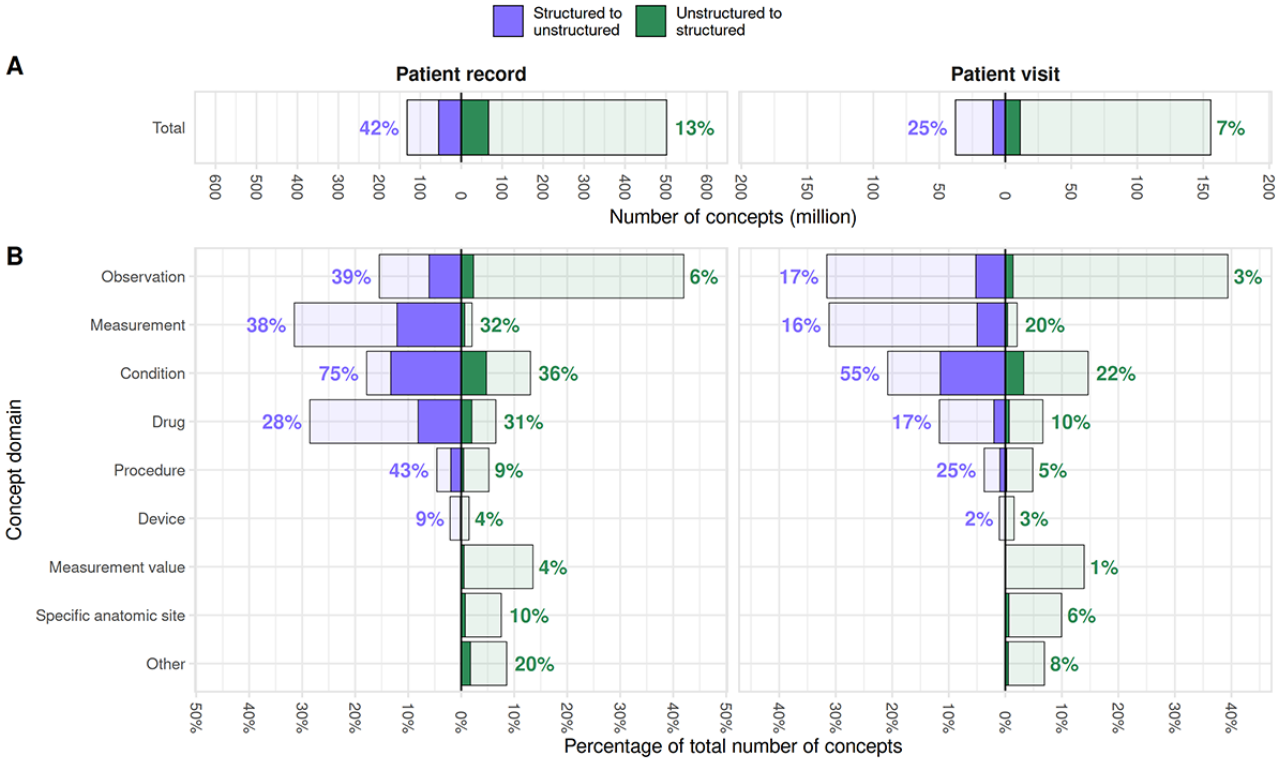


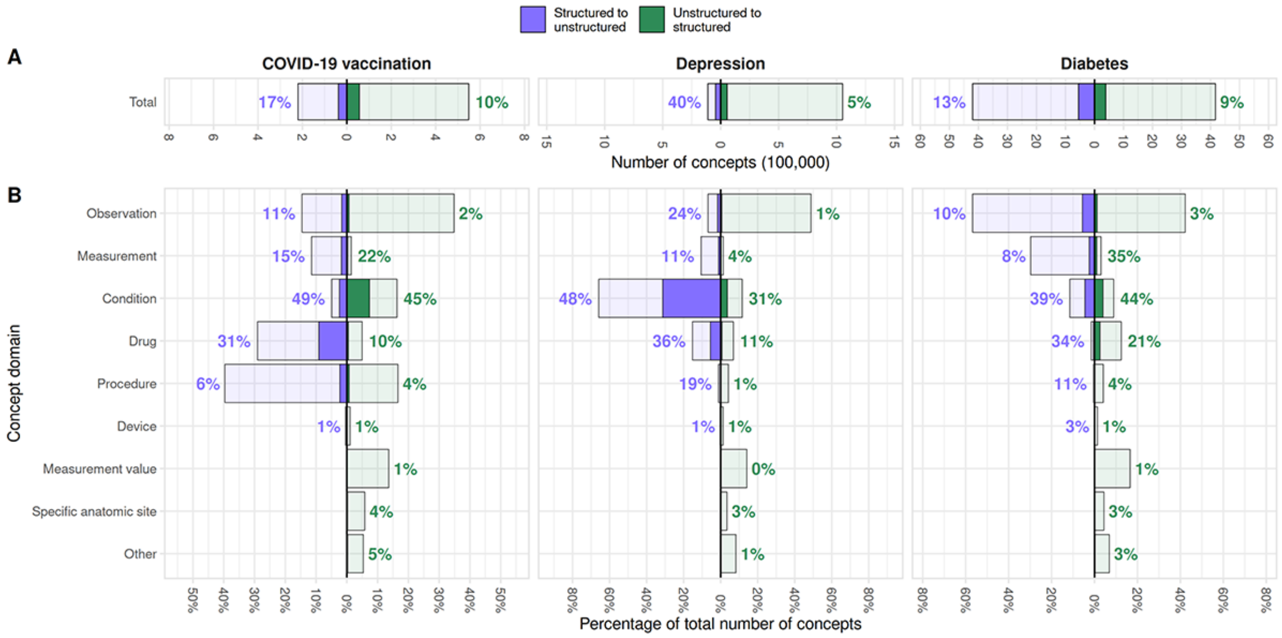
Figure 3B presents the counts and overlap percentages for different domains of structured and extracted concepts. Primarily structured conditions are often also in the text, with 75% (17.6 million of 23.6 million) for patient records and 55% (4.3 million of 7.8 million) for visits. Similarly, extracted condition concepts are also most often matched with structured concepts, with 36% (23.9 million of 65.6 million) in patient records and 22% (5.1 million of 22.8 million) in visits. Other concept domains, such as measurements and drugs, show a relatively high overlap in both structured and unstructured data as well.

Interestingly, differences in concept domain numbers between patient visits and records are larger for structured concepts than extracted concepts. For example, the proportion of structured observation concepts is 31% (11.9 million of 37.8 million) in patient visits but 15% (20.4 million of 132.1 million) in patient records, and the reverse is true for drug concepts. However, this difference is not observed in extracted concepts.

## Differences Between Subpopulations

Figure 4A presents the total concept overlap at the selected threshold for the 3 subpopulations, showing ratios of structured and extracted concepts as seen in Table 1. While the proportion of structured data overlapping with unstructured data and vice versa is similar for the vaccination and diabetes populations, the depression population shows a much higher proportion of structured data matched to concepts in the text (43,934/110,317, 40%) and a lower proportion of extracted concepts matched to structured concepts (55,921/1,051,186, 5%). Figure 4B shows different proportions of concept domains in the structured data across the 3 populations. However, in the extracted concepts, the proportions are relatively similar. The data similarity results are consistent with the total population, with conditions, drugs, and measurement concepts showing the highest overlap between structured and unstructured data.

**Figure 4.** Comparison of structured and extracted concepts in the 3 subpopulation visits. (A) Total number of structured (left) and extracted concepts (right), along with the number of concepts matched at a similarity threshold of 0.65 (blue for structured to unstructured data, green for unstructured to structured). The percentage of matched concepts is listed in the chart for each data type. (B) The proportion of each concept domain contributed to the total number of structured (left) and extracted concepts (right), along with the proportion of concepts in each domain matched to the other data type.



## Discussion

### Comparing Structured and Unstructured Clinical Data

This study explored the feasibility of quantifying the information difference between structured and unstructured data in a large primary care EHR database. We used concept embeddings to measure the similarity between structured clinical event concepts and concepts extracted from free-text narratives. By systematically determining a similarity threshold, we found that a substantial proportion of structured information is also present in unstructured data, while only a small portion of unstructured information is reflected in structured data. This indicates that most concepts in one data type do not match those in the other, suggesting that the information in structured and unstructured data is highly complementary. The difference between the data modalities can be attributed to certain types of information being exclusively structured, such as measurements made at the general practice, while other information, like observations reported in unstructured communication of a hospital to a GP, is often not captured in the structured data of the GP record.

Furthermore, we observed that condition concepts had the largest overlap between structured and unstructured data, followed by measurements and drug concepts. These results were consistent across different data point grouping methods, by patient record or visit. We also quantified the information difference in smaller subpopulations of patients with specific diseases or procedures. Differences in overlap proportions were evident between these subpopulations, but the concept domains with the largest overlap were the same as in the full population. Overall, our findings validate the often assumed notion that unstructured data contains more extensive information than structured data in this specific database. Most importantly, we prove the feasibility of quantifying the information difference in an EHR database by using a combination of clinical concept extraction methods [29], high-dimensional clinical concept embeddings for similarity measurement [22], and annotating a small number of matched concepts to determine the appropriate similarity threshold.

### Strengths and Limitations

The study was conducted on a large, private dataset, which limits public reproducibility. However, our approach to compare and quantify differences between structured and unstructured data is dataset-agnostic. With the detailed pipeline and methodology, the work is adaptable and replicable to any EHR dataset containing both data types. There are several limitations to our methodology. The results depend heavily on 3 factors: the type and quality of concept extraction, the type of concept embeddings used for determining concept similarity, and the chosen similarity threshold. First, the performance of named entity recognition and entity linking is crucial for extracting all relevant information from unstructured data. Imperfect extraction can lead to missed concepts (false negatives) that cannot be matched to structured concepts and wrongly extracted concepts (false positives) that distort the results. The MedSpacy extraction framework used in this work was validated for different languages and demonstrates good performance [29,30]. It was chosen for its availability in Dutch, its versatility with ontologies, and its extraction speed. Second, different types of word and concept embeddings may vary in their ability to distinguish between similar and nonsimilar clinical concepts, affecting the distribution of similarity scores. While other, more complex embeddings might improve similarity discrimination, we only tested the multilingual BioLord-2023-M embedding model [22], as it was specifically trained to produce meaningful representations for clinical sentences and biomedical concepts across different languages, making it perfectly suited to our application. Third, the chosen similarity threshold impacts the results. Setting this threshold can be subjective and heavily depends on the concept embeddings. Therefore, we used a systematic approach using the minimization of the Gini impurity to determine the threshold, and we published the annotated concept matches for transparency. Lastly, data grouping and the ordering of data points in time pose a challenge. We used 2 grouping methods that ignored the time aspect within each group to explore the differences. However, more sophisticated approaches might be necessary depending on the research question. For instance, data recorded outside patient visits may be assigned to the nearest visit, or more advanced sliding time windows could be used to group the data points.

Overall, the strength of this study lies in enhancing our understanding of the information overlap between structured and unstructured clinical data within a large GP EHR database covering 1.8 million patients. Our methodology of using concept embeddings to calculate the similarity between clinical concepts offers a versatile and language-independent solution, demonstrated here for Dutch, ensuring accurate comparisons by capturing nuances in clinical terminology.

### Future Work

Future research in quantifying the data similarity or difference between different data types could explore more advanced concept extraction and embedding techniques or alternative similarity measurements beyond cosine similarity. Our study focused on individual concepts for comparison, but combining multiple concepts might yield higher similarity matches. Investigating the use of concept n-grams and further incorporating context modifiers of the extracted concepts could be beneficial. We applied concept extraction before embedding to retain the granular meaning of individual events within a document. This approach ensures that specific events—such as symptoms, prescriptions, and procedures—are not lost in broader document or sentence embeddings. Future research could explore directly comparing structured concepts and free text without prior clinical concept extraction to enhance the method's reliability and applicability.

We used a single similarity threshold across all concept domains and populations. Future work could involve establishing separate thresholds for each concept domain and population for a more refined comparison. Additionally, considering distinct thresholds for Dutch-to-Dutch and Dutch-to-English concept comparisons might be beneficial, as cross-lingual similarities may vary. The potential applications of our methodology extend beyond the clinical domain. With the appropriate information extraction framework and embedding model, our approach could be adapted to various settings. Conducting specific clinical case studies to demonstrate the benefits of leveraging the data-type differences would be a logical next step for research.

## Conclusions

Our study aimed to assess the feasibility of quantifying the information difference between structured and unstructured data in EHR databases. Within a large Dutch primary care EHR database, we successfully demonstrated that unstructured data provides more extensive information than structured data, quantitatively validating this prevailing assumption. By leveraging concept embeddings to measure semantic similarity between structured concepts and those extracted from free-text narratives, we found that a significant portion of structured information is present in unstructured data, while the reverse occurs much less frequently. Notably, concept domains such as conditions, measurements, and drugs exhibited the largest overlap. Despite limitations related to the performance of concept extraction, the type of embeddings used, and the determination of similarity thresholds, our methodology is versatile and was applied across different data grouping methods and subpopulations. The exploration of more sophisticated techniques could further enhance the accuracy and applicability of this approach. We suggest that structured and unstructured data should be used together, as their combined information exceeds that of each data type separately. Understanding the extent and nature of information in structured and unstructured data within a database can enhance study design, research exploration, resource allocation, and data prioritization, ultimately leading to more robust and insightful observational clinical research.

## Acknowledgments

This work has received support from the European Health Data & Evidence Network (EHDEN) project. EHDEN has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement number 806968. The JU receives support from the European Union’s Horizon 2020 research and innovation program and EFPIA.

## Authors' Contributions

TMS proposed the methodology, designed and implemented the study protocol, and performed the data analysis. JAK, EMvM, and PRR provided critical feedback, helped interpret the results, and shaped the research and analysis. TMS wrote the article with valuable input from all other authors.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Supplementary material.
[DOCX File , 34 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Table S3: annotated concept matches.
[ZIP File (Zip Archive), 98 KB-Multimedia Appendix 2]

## References

1. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis. Mar 2023;82(3):306-311. [FREE Full text] [doi: 10.1136/ard-2022-222626] [Medline: 36150748]
2. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. Aug 01, 2018;25(8):969-975. [FREE Full text] [doi: 10.1093/jamia/ocy032] [Medline: 29718407]
3. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: leveraging the unstructured content in electronic health records. Front Med (Lausanne). 2019;6:66. [FREE Full text] [doi: 10.3389/fmed.2019.00066] [Medline: 31058150]
4. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev. Nov 2022;46:100511. [doi: 10.1016/j.cosrev.2022.100511]
5. Murdoch TB, Detsky AS. The inevitable application of big data to health care. J Am Med Assoc. Apr 03, 2013;309(13):1351-1352. [doi: 10.1001/jama.2013.393] [Medline: 23549579]
6. Chiu CC, Wu CM, Chien TN, Kao LJ, Li C, Chu CM. Integrating structured and unstructured EHR data for predicting mortality by machine learning and latent Dirichlet allocation method. Int J Environ Res Public Health. Feb 28, 2023;20(5):4340. [FREE Full text] [doi: 10.3390/ijerph20054340] [Medline: 36901354]
7. Hashir M, Sawhney R. Towards unstructured mortality prediction with free-text clinical notes. J Biomed Inform. Aug 2020;108:103489. [FREE Full text] [doi: 10.1016/j.jbi.2020.103489] [Medline: 32592755]
8. Kong HJ. Managing unstructured big data in healthcare system. Healthc Inform Res. Jan 2019;25(1):1-2. [FREE Full text] [doi: 10.4258/hir.2019.25.1.1] [Medline: 30788175]
9. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc. Jun 14, 2022;29(7):1292-1302. [FREE Full text] [doi: 10.1093/jamia/ocac058] [Medline: 35475536]
10. Seinen TM, Kors JA, van Mulligen EM, Fridgeirsson E, Rijnbeek PR. The added value of text from Dutch general practitioner notes in predictive modeling. J Am Med Inform Assoc. Nov 17, 2023;30(12):1973-1984. [FREE Full text] [doi: 10.1093/jamia/ocad160] [Medline: 37587084]
11. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc. Aug 2018;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
12. Zhang DD, Yin CC, Zeng JC, Yuan XH, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak. Oct 29, 2020;20(1):280. [FREE Full text] [doi: 10.1186/s12911-020-01297-6] [Medline: 33121479]
13. Gan S, Kim C, Chang J, Lee DY, Park RW. Enhancing readmission prediction models by integrating insights from home healthcare notes: retrospective cohort study. Int J Nurs Stud. Oct 2024;158:104850. [doi: 10.1016/j.ijnurstu.2024.104850] [Medline: 39024965]
14. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. JAMA Netw Open. Dec 07, 2018;1(8):e185097. [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.5097] [Medline: 30646310]
15. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open. Jan 2021;4(1):ooab011. [FREE Full text] [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
16. Park J, Kim K, Hwang W, Lee D. Concept embedding to measure semantic relatedness for biomedical information ontologies. J Biomed Inform. Jun 2019;94:103182. [FREE Full text] [doi: 10.1016/j.jbi.2019.103182] [Medline: 31009761]
17. Zhang Y, Wang X, Lai S, He S, Liu K, Zhao J. Ontology matching with word embeddings. Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer; 2014. Presented at: 13th China National Conference, CCL 2014, and Second International Symposium, NLP-NABD 2014, Wuhan; 2014 October 18-19; China. [doi: 10.1007/978-3-319-12277-9]
18. Abdulnazar A, Kreuzthaler M, Roller R, Schulz S. SapBERT-based medical concept normalization using SNOMED CT. In: Caring is Sharing? Exploiting the Value in Data for Health and Innovation. USA. IOS Press; 2023:825-826.
19. Zahra FA, Kate RJ. Obtaining clinical term embeddings from SNOMED CT ontology. J Biomed Inform. Jan 2024;149:104560. [FREE Full text] [doi: 10.1016/j.jbi.2023.104560] [Medline: 38070816]
20. Xu D, Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. J Biomed Inform. Jun 2022;130:104080. [FREE Full text] [doi: 10.1016/j.jbi.2022.104080] [Medline: 35472514]

21. Liu F, Vuli I, Korhonen A, Collier N. Learning domain-specialised representations for cross-lingual biomedical entity linking. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics; 2021. Presented at: The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; Aug:565-574; Online. [doi: 10.18653/v1/2021.acl-short.72]
22. Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. J Am Med Inform Assoc. Sep 01, 2024;31(9):1844-1855. [doi: 10.1093/jamia/ocae029] [Medline: 38412333]
23. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. JMIR Med Inform. Aug 02, 2019;7(3):e13802. [FREE Full text] [doi: 10.2196/13802] [Medline: 31376277]
24. Bucher BT, Shi J, Pettit RJ, Ferraro J, Chapman WW, Gundlapalli A. Determination of marital status of patients from structured and unstructured electronic healthcare data. AMIA Annu Symp Proc. 2019;2020:267-274. [FREE Full text] [Medline: 32308819]
25. Ruckdeschel JC, Riley M, Parsatharathy S, Chamarthi R, Rajagopal C, Hsu HS, et al. Unstructured data are superior to structured data for eliciting quantitative smoking history from the electronic health record. JCO Clin Cancer Inform. Feb 2023;7:e2200155. [doi: 10.1200/cci.22.00155]
26. Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. Cancer Inform. Dec 08, 2016;15:CIN.S40604. [doi: 10.4137/cin.s40604]
27. de Ridder MA, de Wilde M, de Ben C, Leyba AR, Mosseveld BM, Verhamme KM, et al. Data resource profile: the Integrated Primary Care Information (IPCI) database, the Netherlands. Int J Epidemiol. 2022;51(6):e314-e323. [FREE Full text] [doi: 10.1093/ije/dyac026] [Medline: 35182144]
28. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574-578. [FREE Full text] [doi: 10.3233/978-1-61499-564-7-574] [Medline: 26262116]
29. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438-447. [FREE Full text] [Medline: 35308962]
30. Seinen TM, Kors JA, van Mulligen EM, Rijnbeek PR. Annotation-preserving machine translation of English corpora to validate Dutch clinical concept extraction tools. J Am Med Inform Assoc. Aug 01, 2024;31(8):1725-1734. [doi: 10.1093/jamia/ocae159] [Medline: 38934643]
31. SNOMED National Release Centre of the Netherlands 2024. URL: https://www.snomed.org/member/netherlands [accessed 2024-09-25]
32. Afzal Z, Pons E, Kang N, Sturkenboom MC, Schuemie MJ, Kors JA. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. BMC Bioinformatics. Nov 29, 2014;15(1):373. [FREE Full text] [doi: 10.1186/s12859-014-0373-3] [Medline: 25432799]
33. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711-3718. [FREE Full text] [doi: 10.1093/bioinformatics/bty373] [Medline: 29757357]

## Abbreviations

**EHR:** electronic health record
**GP:** general practitioner
**IPCI:** Integrated Primary Care Information
**UMLS:** unified medical language system

*Edited by G Tsafnat; submitted 04.10.24; peer-reviewed by F Remy, B Van Es, Y Liu; comments to author 02.11.24; revised version received 22.11.24; accepted 23.11.24; published 13.02.25*

*Please cite as:*
*Seinen TM, Kors JA, van Mulligen EM, Rijnbeek PR*
*Using Structured Codes and Free-Text Notes to Measure Information Complementarity in Electronic Health Records: Feasibility and Validation Study*
*J Med Internet Res 2025;27:e66910*
*URL: https://www.jmir.org/2025/1/e66910*
*doi: 10.2196/66910*
*PMID:*