Measuring Redundancy in Czech Electronic Health Records: Near-Duplicate Detection and Cluster Analysis

Kristof Anetta and Aleš Horák

Natural Language Processing Centre, Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic
xanetta@fi.muni.cz, hales@fi.muni.cz

Abstract. Electronic health records (EHRs) contain extensive repetition arising from templated structures, copy-paste practices, and recurrent clinical phrasing. While such redundancy facilitates documentation consistency, it also affects the efficiency of data processing and downstream natural language processing applications.
This study investigates the internal textual redundancy of a Czech dataset of narrative parts of oncology health records using a fast near-duplicate detection method and a subsequent clustering analysis. We quantify the degree and distribution of repeated content across documents, visualize the resulting clusters to identify patterns, and experiment with creating cluster-aware pruned datasets for more efficient language model training. For comparison, we report baseline redundancy measures on a Czech literary corpus, illustrating the contrast between natural and clinical text. In addition to providing insight into how redundancy shapes the linguistic and informational landscape of Czech EHRs, we discuss our findings in the context of state-of-the-art clinical LLMs for English, making a case not only for continued development of redundancy-mitigating approaches, but also for the use of synthetic health record data.

Keywords: Electronic health records; EHR; corpus; dataset; redundancy; near-duplicate; deduplication; Czech.

1 Introduction

When processing electronic health records for the training of large language models, redundancy is a significant consideration; substantial portions of clinical documentation consist of repeated or minimally altered formulations, which inflate corpus size without proportionally increasing linguistic or informational diversity. This issue has been addressed many times in research literature.

An individual researcher manually inspecting tens or hundreds of records usually senses the repetition but rarely glimpses the true scale of it across tens of thousands to millions, and therefore has a limited capacity to judge to what extent this redundancy may constrain the generalization of the future model,

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 117–125, 2025.

making it effectively encounter the same expressions of clinical events rather than diverse realizations of underlying concepts.

Exploring the nature and extent of this phenomenon on a large dataset of real clinical notes, this study analyzes how textual redundancy manifests in Czech oncology health records. By examining this dataset through near-duplicate detection and clustering methods, we aim to characterize the forms and scales of redundancy that arise in this specific language and institutional environment and to assess their implications for corpus design and model training.

2 Related work

Quantitative studies have shown that clinical notes often contain extensive duplicated content – on average exceeding 50% in progress and signout notes – and that note length and redundancy have both increased significantly over the past decade across specialties and institutions [11,15,16]. Such repetition not only inflates record size but also propagates inconsistencies and hampers the usability of health records. Cross-linguistic comparison further indicates that redundancy levels vary culturally or institutionally: French EHRs exhibit substantially lower duplication than U.S. corpora, likely reflecting different documentation norms [5].

Complementary research has focused on measuring and mitigating this redundancy. Alignment-based and lexical-overlap approaches have been validated against expert ratings as reliable estimators of duplicated information, while information-theoretic and transformer-based methods have quantified the informational inefficiency of clinical text relative to open-domain corpora [12,15,16]. To address redundancy's impact on NLP, redundancy-aware preprocessing and modeling frameworks – such as fingerprinting techniques and Red-LDA topic modeling – have been proposed, yielding more coherent topics and improved text-mining performance [3,4]. Collectively, these studies establish redundancy as both a measurable structural feature of EHR corpora and a critical factor influencing clinical NLP effectiveness.

3 Data

The EHR data used in this study are oncology health records collected at the Masaryk Memorial Cancer Institute in Brno, Czech Republic. This unannotated, 42-million-word dataset of clinical documentation, containing 154,763 health records, referred to as CSEHR_corpus, is the subject of ongoing research. In previous experiments, we have created an annotated subset of this data, named CSEHR [1].

4 Deduplication analysis: EHRs vs. literary corpus

As a first exploratory step in understanding redundancy in Czech EHRs, we utilized the Onion corpus deduplication tool [10]. By default, Onion deduplicates

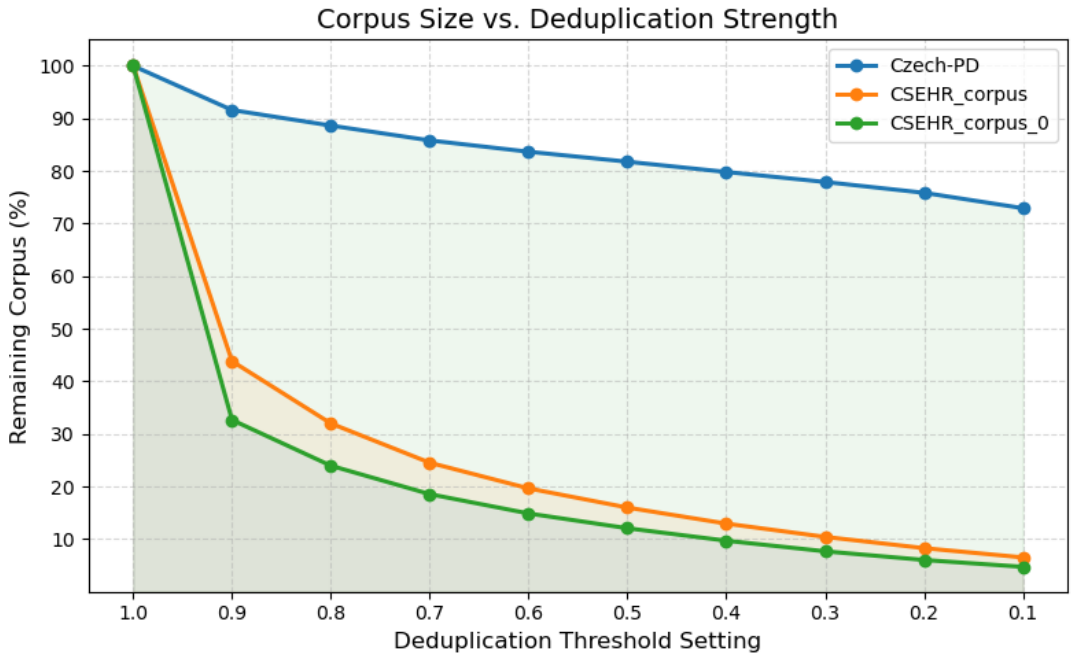


Fig. 1: Visualization of the redundancy gradient of EHR data compared to the baseline of Czech-PD [9] (public domain literary works), expressed in the progressive decline of corpus size as Onion [10] deduplication strength is increased.

using 5-gram hashes trimmed to 64 bits, and maximum stub length of 20. We set it to remove duplicate sentences (the `-p` parameter set to `s`), testing a range of values of the duplicate content threshold (from 0.1 to 0.9 in increments of 0.1). We applied the tool to three different datasets:

1. Czech-PD, a corpus of public domain literary works [9]
2. CSEHR_corpus
3. CSEHR_corpus_0 (all numeric values in texts replaced by 0 to diminish the variety introduced by different measurements, dates, identifiers, etc.)

Czech-PD was included as a control dataset and it serves as a rough baseline of how a non-deduplicated dataset of natural language responds to different strengths of deduplication.

The results are displayed in Figure 1. Visually, the most dramatic drop occurs between full dataset size and deduplication threshold of 0.9, where 92% is left of Czech-PD, but EHRs have been reduced to 44%, in the case of EHRs with zeroes 33%. Numerically, though, the biggest disproportion is still to come. At the default setting of 0.5, 82% is left of Czech-PD, but EHRs have sunk to 16%, zeroed ones to 12%. At the most severe threshold of 0.1, Czech-PD retains 73% of all data, but EHRs are at 7%, zeroed EHRs at 5%.

These results show that redundancy in clinical text operates on multiple levels of similarity, extending far beyond exact duplication. The steep reduction

between the full corpus and the 0.9 threshold indicates that a large proportion of documents are either fully identical or differ only trivially, for example by patient identifiers or minor lexical edits. Yet the continued, disproportionate decline at lower thresholds reveals an additional layer of near-duplicates and paraphrastic variants: records that share large textual segments or templated phrasing while differing in surface details. In contrast, the Czech-PD corpus remains comparatively stable across thresholds, suggesting that natural language redundancy is largely limited to stylistic or collocational repetition. The cascading losses visible in Figure 1 thus illustrate how EHRs are structured around recurring textual scaffolds, where both identical and closely similar records dominate the dataset.

5 Clustering

To explore internal textual structure and identify groups of near-identical documents, we employed clustering based on MinHash [2] and Locality-Sensitive Hashing (LSH) [7]. These methods are designed to efficiently estimate pairwise similarity between large numbers of documents without computing full edit distances or cosine similarities on high-dimensional vectors.

MinHash provides an approximation of the Jaccard similarity between two sets of tokens or shingles by representing each document through a compact signature of hash minima. Documents sharing many overlapping n-grams have similar signatures, which allows their similarity to be estimated probabilistically. Building on this, Locality-Sensitive Hashing partitions the signature space into hash buckets such that similar documents are likely to fall into the same bucket. This greatly reduces the computational cost of pairwise comparisons and enables large-scale clustering of near-duplicate or templated texts.

In this study, each document was represented as a set of 5-gram shingles. We computed MinHash signatures using 256 hash functions and indexed them with LSH to obtain candidate groups of highly similar records. These groups were then refined through average linkage clustering on the estimated Jaccard distances, yielding clusters that correspond to recurring templates, partially reused sections, or shared report structures within the corpus.

5.1 Visualizing superclusters

In order to make the similarities between individual documents as palpable as possible, we

- ran a greedy clustering algorithm with a Jaccard threshold of 0.5
- extracted all clusters with more than 100 documents, totaling 58 clusters containing 27,039 of all 154,763 documents (17.47%)
- arranged the clusters based on cluster-to-cluster centroid similarities
- and, reducing clusters to ca. 10% of their respective sizes, plotted the 2,681 x 2,681 document similarity heatmap in Figure 2.

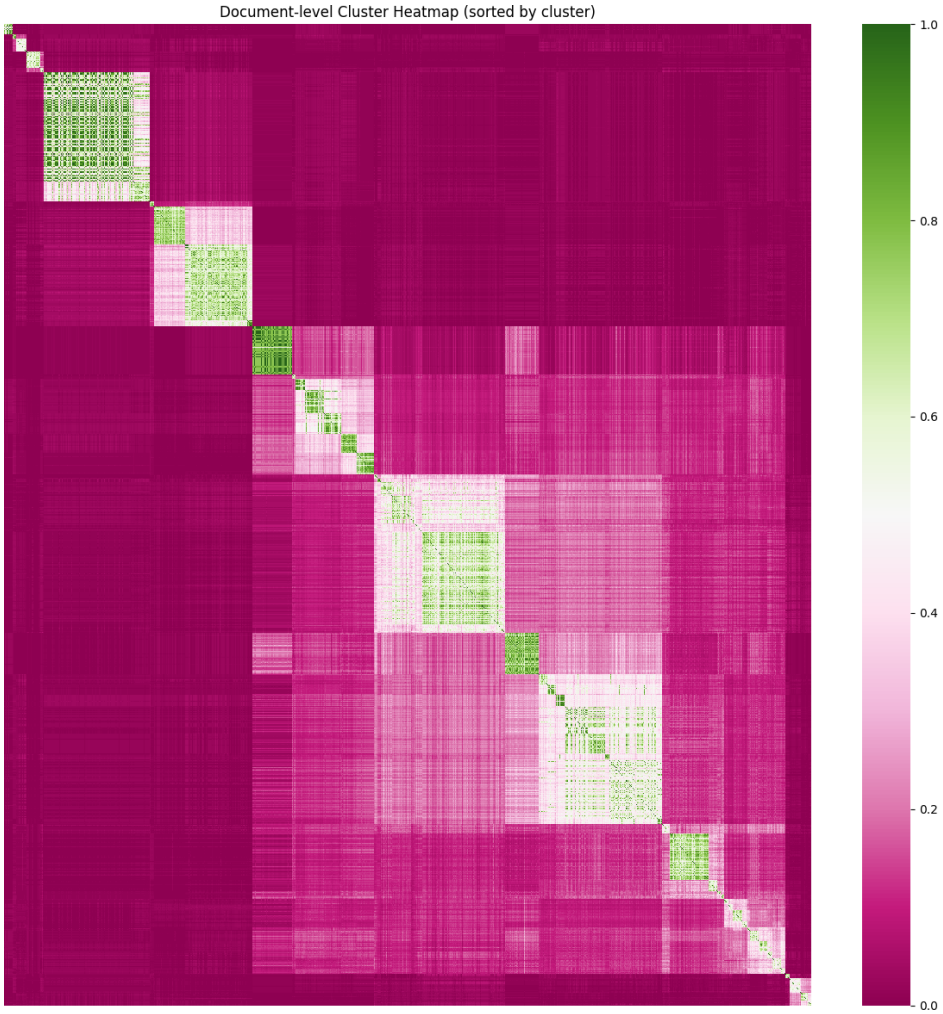
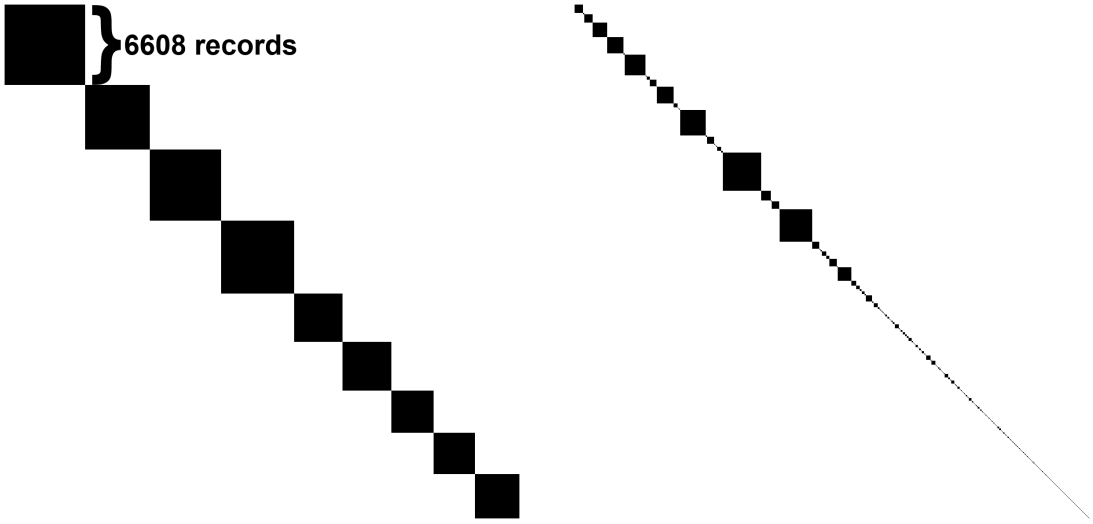


Fig. 2: A heatmap of the Jaccard similarities of documents in all clusters with more than 100 documents (top 58 clusters). Clusters were reduced to 10% of their original size for plotting. White color marks similarity around the clustering threshold of 0.5. Green spectrum signifies similarity increasing past 0.5 up to 1.0, red spectrum traces how more subtle pink similarities under 0.5 fade out to 0.

This visualization is very good at showing the pervasive patterns of redundancy in EHR datasets. Beyond the greenish cluster "beads" along the diagonal, marking almost identical records, we can see pink rectangles of increasing size that signify "superclusters", expanding the core clusters into wider and wider networks of similarity and interrelation.

Figure 3 shows the difference in granularity when visualizing the most clustered 25% of CSEHR_corpus with different redundancy thresholds.



(a) (b)

Fig. 3: Binary heatmaps showing the most clustered 25% of CSEHR_corpus, comparing the cluster structure with redundancy threshold 0.2 (a) and 0.5 (b).

5.2 Cluster-aware dataset pruning

As an alternative to brute force sentence/paragraph/document-level deduplication offered by Onion [10], we carried out cluster-aware dataset pruning in several different configurations. By keeping only 1 record from each cluster, the resulting pruned datasets we experimented with included:

- CSEHR_corpus_cluster50: Pruned using clusters with threshold 0.5, reduced to 82,059 documents and 56.7% of original size
- CSEHR_corpus_cluster20: Pruned using clusters with threshold 0.2, reduced to 17,301 documents and 8.3% of original size.

Building on simple pruning, future possibilities in leveraging clusters include more complex approaches to clinical language model training, where the diversifying benefits of pruning are maintained and total size remains efficient, but the intra-cluster variety, lost in simple pruning, is leveraged (e.g. by switching cluster members in training data between epochs).

6 Discussion: A case for synthesis?

By showing the gradient of redundancy and calculating similarity clusters among EHRs, our study emphasises the degree to which redundancy serves as a fundamental organizing principle of clinical documentation.

The implication: when training large language models, corpora of clinical text might need to receive treatment different from that of most natural language

corpora, where redundancy is mostly addressed in the step of near-duplicate removal. Without addressing the deeper structural redundancy of clinical text, the performance of language models trained on it is likely to suffer proportionally from the lack of variety.

This begs the question what options there are for further processing. A natural first choice is "extended deduplication" whereby the criteria for exclusion are changed. This might include preprocessing text to remove variety introduced by numbers or codes if language stays the same, or a reduction of the size of compared chunks, so that copy-pasted lines can be excluded on their own and do not hide in a larger document.

But so far, these have all been ways to reduce data size, which might be good for efficiency, but not a primary way of increasing performance.

State-of-the-art research in large-scale clinical language modeling offers an interesting case: in the GatorTron family of models, GatorTronS trained on 1B, 5B, 10B, and 20B tokens of synthetic EHRs generated by GatorTronGPT has generally outperformed GatorTron trained on 90B tokens of real EHR data [8]. Even the 1B synthetic data model has matched the results of the much larger 90B real data model, and the 5B model has already outperformed it in clinical concept extraction and medical relation extraction benchmarks (2010 i2b2 [14], 2012 i2b2 [13], and 2018 n2c2 [6]).

One plausible explanation lies in redundancy and diversity of training data. The authors compared synthetic data with real data and found synthetic text significantly superior in the diversity of unigrams (40.43M vs. 4.82M), the diversity of bigrams (416.35M vs. 62.51M) and text entropy (4.97 vs. 4.95). In short, synthetic datasets seem to successfully mitigate the limitations of available real data by introducing lexical and structural variation while preserving domain semantics, improving model generalization.

This advances the discussion from the question "How to best process EHR datasets to create clinical text corpora?" to "How to achieve the best synergy between optimal processing of real EHR datasets and generating synthetic data when creating clinical text corpora?", which is a reasonable next step in the Czech language space and beyond.

7 Conclusion

This study provided the first large-scale quantitative analysis of textual redundancy in Czech electronic health records. Using near-duplicate detection and MinHash–LSH clustering, we demonstrated that repetition in Czech oncology documentation is pervasive, multi-layered, and structurally embedded, extending far beyond exact copies to include wide networks of templated and partially reused formulations. Comparison with a literary corpus highlighted how strongly EHRs deviate from typical natural language in their redundancy profiles, underscoring the need for domain-specific preprocessing strategies.

Addressing this redundancy requires methods that respect the structural regularities of documentation while expanding linguistic diversity, whether

through refined preprocessing, template-aware reduction, or controlled generation of synthetic data. Future work should explore this balance more explicitly, integrating redundancy-aware corpus design with generative augmentation to maximize the representational and inferential value of clinical corpora, serving the next generation of clinical NLP systems in Czech and other under-resourced languages.

Acknowledgments. The analyzed Czech data was collected at the Masaryk Memorial Cancer Institute in Brno. Research has been partly supported by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101079183 (BioMedAI). This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062. Computational resources were provided by the e-INFRA CZ project ID:90254 supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Anetta, K., Horák, A.: New human-annotated dataset of czech health records for training medical concept recognition models. In: International Conference on Text, Speech, and Dialogue. pp. 110–120. Springer (2024)
2. Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171). pp. 21–29. IEEE (1997)
3. Cohen, R., Aviram, I., Elhadad, M., Elhadad, N.: Redundancy-aware topic modeling for patient record notes. PloS one **9**(2), e87555 (2014)
4. Cohen, R., Elhadad, M., Elhadad, N.: Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC bioinformatics **14**(1), 10 (2013)
5. D'hondt, E., Tannier, X., Névéol, A.: Redundancy in french electronic health records: A preliminary study. In: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis. pp. 21–30 (2015)
6. Henry, S., Buchan, K., Filannino, M., Stubbs, A., Uzuner, O.: 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association **27**(1), 3–12 (2020)
7. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing. pp. 604–613 (1998)
8. Peng, C., Yang, X., Chen, A., Smith, K.E., PourNejatian, N., Costa, A.B., Martin, C., Flores, M.G., Zhang, Y., Magoc, T., et al.: A study of generative large language model for medical research and healthcare. NPJ digital medicine **6**(1), 210 (2023)
9. PleIAs: Czech-PD: Czech public domain corpus. `https://huggingface.co/datasets/PleIAs/Czech-PD` (2024), accessed: 2025-11-04
10. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic (2011)
11. Rule, A., Bedrick, S., Chiang, M.F., Hribar, M.R.: Length and redundancy of outpatient progress notes across a decade at an academic medical center. JAMA Network Open **4**(7), e2115334–e2115334 (2021)

12. Searle, T., Ibrahim, Z., Teo, J., Dobson, R.: Estimating redundancy in clinical text. Journal of biomedical informatics **124**, 103938 (2021)
13. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association **20**(5), 806–813 (2013)
14. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (2011)
15. Wrenn, J.O., Stein, D.M., Bakken, S., Stetson, P.D.: Quantifying clinical narrative redundancy in an electronic health record. Journal of the American Medical Informatics Association **17**(1), 49–53 (2010)
16. Zhang, R., Pakhomov, S., McInnes, B.T., Melton, G.B.: Evaluating measures of redundancy in clinical texts. In: AMIA annual symposium proceedings. vol. 2011, p. 1612 (2011)