technologies

MDPI

Article

Comparative Evaluation of Rule-Based and Transformer-Based Text-Mining Methods for Detecting SGLT2 Inhibitor Mentions in Unstructured Clinical Free Text

Attila Csaba Nagy

Department of Epidemiology, Faculty of Health Sciences, University of Debrecen, 4028 Debrecen, Hungary; nagy.attila@etk.unideb.hu

Abstract

Much of the patient data recorded in electronic health records is stored as unstructured free text. Extracting medication information from such data is essential, particularly for antidiabetic drugs such as sodium–glucose cotransporter-2 (SGLT2) inhibitors, but remains challenging due to spelling variability, abbreviations, and non-standard documentation practices. This study compared four text-mining approaches, simple keyword search, regular expression–based matching, fuzzy string matching, and a transformer-based token classification baseline, for detecting SGLT2 inhibitor mentions in Hungarian clinical narratives. Clinical documents were obtained from the University of Debrecen Clinical Centre and covered patients with type 2 diabetes mellitus (ICD-10: E11) from 2018 and 2019. Searches targeted both generic and brand names and SGLT-related abbreviations. In the 2019 dataset (n = 5383), simple keyword search identified 1.49% of documents as containing an SGLT2 inhibitor mention, compared with 7.21% using regular expressions, 8.55% using fuzzy matching, and 0.71% using the transformer-based baseline. Mean execution times were 0.07 s, 1.64 s, 5.13 s, and 34.71 s, respectively. Method performance was further evaluated against a manually annotated reference set from 2018 using confusion matrices and standard classification metrics. Fuzzy string matching achieved the highest recall and F1-score, while regular expression-based matching provided a strong balance between precision and recall. The transformer-based baseline showed high precision but substantially lower recall in the absence of domain-specific fine-tuning. Overall, similarity-based fuzzy matching offered the most favorable balance between detection performance and computational efficiency for identifying SGLT2 inhibitor mentions in unstructured Hungarian clinical text.

Keywords: text mining; fuzzy matching; regular expression; clinical NLP; SGLT2 inhibitors

Academic Editor: Sheryl Berlin Brahnam

Received: 15 January 2026
Revised: 10 February 2026
Accepted: 13 February 2026
Published: 15 February 2026

1. Introduction

In today's world, vast amounts of data are created every day in various areas, including medicine. All features of the traditional "big data" paradigm (value, volume, velocity, variety, veracity and variability) manifest in this field [1,2]. Health data are extremely diverse and polystructural, as they originate from many sources, including laboratory information systems, diagnostic imaging, measurements from wearable sensors, and narrative clinical data. Graphical and computational methods are increasingly applied for visualization and interpretation of complex datasets, allowing clinicians and researchers to monitor longitudinal data and to explore dependencies beyond therapeutic pathways [3,4].

Health information systems (HIS) face significant challenges when processing and analyzing unstructured data. Unstructured data, such as free-text clinical notes, medical reports, images, and PDF documents, represent approximately 80% of all health data [5]. While structured clinical fields support efficient data processing, they often miss relevant contextual information that is frequently documented only in unstructured records. When appropriately processed and analyzed, unstructured data can support data-driven clinical decisions, treatment pattern analysis, precision medicine, tertiary prevention, and improvements in quality of life [5–8]. In clinical narrative documentation, therapeutic rationale, physician preferences, medication changes, and patient-reported symptoms are often poorly captured or entirely absent from structured fields of electronic health record (EHR) systems. As a result, unstructured clinical text has become a core component of contemporary clinical informatics, particularly for large-scale observational studies and real-world evidence generation.

To realize this potential, precise retrieval of therapeutic information from narrative records is required [9]. Most long-term medications for chronic diseases, including type 2 diabetes mellitus, are recorded in free-text clinical narratives rather than in formal prescription tables. These circumstances have led to an increasing demand for reliable text-mining techniques capable of identifying drug names, formulations, and clinical indications while accounting for the noise and heterogeneity of routine clinical documentation. In this regard, extraction of medication history and prescription information is severely hindered by the lack of standardized formats and by institution-specific documentation practices.

Over recent years, healthcare systems have evolved from manually curated databases into extensive digital platforms that generate clinical data in a continuous manner. Structured EHR fields capture relevant diagnostic and therapeutic events, but often provide only a partial representation of the patient's clinical course. Clinicians frequently document the reasoning behind treatment decisions, dosage adjustments, and nuanced clinical assessments exclusively in free-text fields. Consequently, narrative documentation plays a key role in understanding real-world clinical practice. In this context, text-mining techniques have been established as effective tools for extracting valuable information from clinical EHR data. This is particularly relevant for research areas requiring precise exposure assessment, such as pharmacoepidemiology, cardiovascular primary prevention, and chronic disease research [10,11].

Extraction of medication information from clinical text is among the most challenging tasks in clinical natural language processing. Ambiguities arising from inconsistent spelling, code mixing, informal abbreviations, and substantial inter-clinician variation are further amplified in morphologically rich languages such as Hungarian. The inflectional morphology of Hungarian produces multiple surface forms for a single drug concept, posing a challenge for string-matching-based NLP approaches. As a result, medication mentions are frequently missed, and drug exposure may be systematically under-reported in retrospective text-based studies. Although rule-based text-mining methods and transformer-based language models have been proposed to address these challenges, their comparative evaluation in Hungarian clinical text remains limited [12,13].

Sodium–glucose cotransporter-2 (SGLT2) inhibitors are a class of oral antidiabetic drugs that provide glycemic control along with cardiovascular and renal protective effects [6,14–17]. Given their increasing clinical use and relevance to treatment guidelines, it is important to assess how SGLT2 inhibitors are captured in real-world clinical data. Accurate assessment of SGLT2 inhibitor use requires text-mining methodologies capable of detecting mentions of this drug class in unstructured clinical text, including instances involving spelling errors, abbreviations, or linguistic variation.

The challenge is further compounded by variability in narrative documentation. Clinicians often use shortened brand names, acronyms, non-standard multilingual variants, or phonetic spellings, particularly under time constraints. Previous studies suggest that up to one third of medication mentions may occur in non-standard formats, substantially reducing the performance of automatic extraction systems [12]. Such documentation practices may introduce bias into pharmacoepidemiological analyses, particularly when informal documentation styles are unevenly distributed across clinical settings or patient groups.

Morphological complexity is especially pronounced in agglutinative languages such as Hungarian, where grammatical context is expressed through inflection and compounding of medication names (e.g., *dapagliflozinnal*, *dapagliflozint*). Prior work in multilingual NLP has shown that morphologically rich languages often require pattern-based or similarity-based approaches to achieve adequate recall [18]. In addition, Hungarian clinical documents frequently exhibit mixed English and Hungarian spelling conventions, particularly in brand names and combination therapies. These localized and institution-specific variants further necessitate flexible and adaptive extraction strategies.

From a clinical and epidemiological perspective, reliable identification of SGLT2 inhibitor use is essential. Although international uptake of these agents is increasing, substantial heterogeneity remains across healthcare systems [19]. Accurate detection of medication mentions in unstructured clinical narratives enables real-world effectiveness research, guideline adherence assessment, safety evaluation, and quality-of-care studies. Improved extraction performance also supports downstream analyses of treatment pathways and patient stratification.

Existing text-mining approaches for medication detection can be broadly categorized as rule-based, similarity-based, or machine-learning-based methods [20]. Rule-based approaches, such as keyword search and regular expressions, are transparent and computationally efficient but sensitive to lexical variation. Similarity-based techniques, including fuzzy string matching, are more robust to spelling and morphological variability. Machine-learning approaches, often based on transformer architectures, capture contextual information but are typically resource-intensive and require domain-specific training data. While hybrid methods combining these approaches have been reported to improve performance [20,21], systematic evaluations in Hungarian clinical text remain scarce.

Accordingly, the primary objective of this study was to compare multiple text-mining approaches in terms of precision, computational efficiency, and practical applicability for identifying SGLT2 inhibitor mentions in unstructured Hungarian clinical narratives.

2. Materials and Methods

The presence of type 2 diabetes mellitus (T2DM) was confirmed using the ICD-10 diagnostic code E11. Administrative and clinical data were obtained from the University of Debrecen Clinical Centre for the period 2007 to 2021. Only data from 2018 and 2019 were included in the present analysis, as these years represent pre-COVID periods with stable clinical activity and minimal administrative disruption.

All analyses were conducted using Python (version 3.12.7; Python Software Foundation, Beaverton, OR, USA) [22], an open-source programming language widely used for scientific computing and data analysis. Detection of SGLT2 inhibitor mentions was based on a curated lexicon of generic and brand names compiled from contemporary therapeutic guidelines. The lexicon included dapagliflozin, Forxiga, Xigduo, empagliflozin, Jardiance, Synjardy, ertugliflozin, Steglatro, Steglujan, and the general abbreviation SGLT.

2.1. Dataset Description

The corpus comprised standard clinical text types, including discharge letters, inpatient flow notes, outpatient referrals, laboratory reports, and emergency department notes. The clinical language of Hungarian medical documentation is highly variable and includes shorthand expressions, codes, mixed Hungarian–English terminology, and inflected or shortened drug names, consistent with previous studies on Hungarian clinical NLP [18,23].

Text normalization included Unicode normalization, white-space normalization, and removal of non-printing control characters. Lemmatization and stemming were deliberately omitted, as over-normalization may distort pharmaceutical tokens in morphologically rich languages such as Hungarian and thereby reduce recall in medication identification tasks [18,24].

2.2. Text-Mining Approaches

The selected approaches were chosen to cover commonly used text-mining strategies with different levels of complexity and computational requirements, enabling a pragmatic comparison across rule-based, similarity-based, and transformer-based methods. Four text-mining approaches were evaluated:

1. Simple keyword search
2. Regular expression-based matching
3. Fuzzy string matching
4. Transformer-based token classification baseline (HuBERT)

2.2.1. Simple Keyword Search

The simple keyword search approach applied direct substring matching using the predefined medication lexicon. While computationally efficient, this method was limited to exact matches and therefore did not account for spelling variation, abbreviations, or morphologically inflected forms common in Hungarian clinical text. These limitations have been widely reported in prior medication extraction studies.

2.2.2. Regular Expression Matching

Regular expression patterns were designed to capture common sources of lexical variation, including non-standard spacing, partial matches, and frequent Hungarian morphological suffixes (e.g., *-ban*, *-val*, *-t*). Although this approach provided greater flexibility than exact keyword matching, it remained sensitive to previously unseen misspellings and non-standard abbreviations, consistent with earlier findings on rule-based clinical extraction systems.

2.2.3. Fuzzy String Matching

Fuzzy string matching was implemented using the RapidFuzz library (version 3.0.0; open-source, GitHub, San Francisco, CA, USA), based on Levenshtein-distance similarity metrics. Similarity thresholds were evaluated in a pilot phase and fixed between 80% and 90% to balance recall and precision. This approach enabled broad coverage of lexical variants, including inflected, abbreviated, and misspelled drug names in Hungarian clinical text. Comparable benefits of similarity-based methods have been reported in previous evaluations involving noisy clinical corpora.

2.2.4. Transformer-Based Token Classification Baseline

A transformer-based token classification model using the HuBERT language model (available via Hugging Face, Inc., New York, NY, USA) was evaluated as a baseline named entity recognition-like approach. The model was pretrained on general Hungarian language

corpora and applied off the shelf, without domain-specific clinical fine-tuning. For practical reasons, token classification was restricted to candidate documents prefiltered by rule-based methods.

Transformer-based classifiers often exhibit reduced recall in clinical text due to domain mismatch, subword tokenization effects, and the low frequency of medication mentions [25,26]. Accordingly, this method was included as a comparative baseline rather than as a fully trained clinical NER system.

2.3. Manual Annotation and Reference Set

For performance evaluation, the 2018 dataset was manually annotated to create a reference validation set. Candidate documents were identified using keyword-assisted screening based on partial lexical similarity to SGLT2-related terms. Each candidate document was reviewed manually by a clinical researcher to confirm true medication mentions.

Negated, historical, or non-current references, such as discontinued therapies or exclusion statements, were classified as negative. False-positive outputs generated by the automated approaches were reconciled against the manual annotations and incorporated into the final reference set.

2.4. Statistical Analysis

Detection proportions were reported with 95% confidence intervals, and Z-tests were used to compare proportions, with statistical significance defined as $p < 0.05$. Descriptive statistics were calculated for document length, including word count and character count. Given the skewed distribution of free-text variables, means with standard deviations were reported alongside medians with interquartile ranges (IQR).

Descriptive statistics were computed in Python using the pandas (version 2.1.0; open-source, NumFOCUS, Austin, TX, USA) and numpy (version 1.26.0; open-source, NumFOCUS, Austin, TX, USA) libraries. Confidence intervals and hypothesis tests were independently verified using Stata (version 18; StataCorp LLC, College Station, TX, USA) [27] to ensure computational accuracy and reproducibility. Normality was assessed using the Shapiro–Wilk test.

3. Results

A total of 11,085 clinical documents were analyzed across the two study years (2018–2019). The 2019 dataset comprised 5383 patients with type 2 diabetes mellitus, while 5704 patients were included in 2018. Across both years, the median document length was 156 words (interquartile range [IQR]: 75–293; mean: 235 words) and 1292 characters (IQR: 573–2446; mean: 1933 characters), with substantial variability (ranges: 0–5360 words; 0–39,568 characters). When stratified by year, the median word count was slightly higher in 2019 (164 words; IQR: 79–302) compared with 2018 (150 words; IQR: 71–283), indicating marginally more detailed documentation in the later period.

Misspellings and orthographic variants of SGLT2 inhibitors, such as forxida, jardianc, and empaglifozin, as well as abbreviations including sglt, sglt2, and flozin, were frequently observed in the clinical narratives.

3.1. Document-Level Detection Rates Across Text-Mining Methods

Document-level detection rates differed substantially across the evaluated text-mining approaches. In the 2019 dataset, simple keyword search identified 80 documents containing an SGLT2 inhibitor mention (1.49%, 95% CI: 1.17–1.81%). Regular expression–based matching detected 388 documents (7.21%, 95% CI: 6.52–7.90%), while fuzzy string matching identified 460 documents (8.55%, 95% CI: 7.80–9.30%). The transformer-based token classification baseline identified 38 documents (0.71%, 95% CI: 0.48–0.93%).

Comparable patterns were observed in the 2018 dataset, with document-level detection rates ranging from 1.28% for simple keyword search to 6.42% for fuzzy string matching. These document-level detection rates are presented for descriptive purposes and provide context for subsequent validation analyses.

3.2. Validation Against the Manually Annotated Reference Set

Performance of the evaluated text-mining approaches was assessed against the manually annotated reference set from 2018 (Table 1). Confusion matrices for each method are shown in Table 2, and corresponding performance metrics are summarized in Table 3.

Table 1. Characteristics of the manually annotated reference set (2018).

Variable	Value
Study year	2018
Total number of clinical documents	5704
Documents with ≥1 confirmed SGLT2 inhibitor mention	323
Documents without SGLT2 inhibitor mention	5381
Total confirmed SGLT2-positive documents (gold standard)	323
Annotation method	Manual expert annotation
Evaluation unit	Document-level (binary)

Table 2. Confusion matrices of SGLT2 inhibitor detection methods against the manually annotated reference set (2018).

Method	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)
Simple keyword search	61	12	262	5369
Regular expression matching	294	18	29	5363
Fuzzy string matching	323	43	0	5338
Transformer-based token classification (HuBERT)	129	3	194	5376

Table 3. Performance metrics of text-mining methods evaluated against the manually annotated reference set (2018).

Method	Precision	Recall	F1-Score	Specificity	Accuracy
Simple keyword search	0.836	0.189	0.308	0.998	0.952
Regular expression matching	0.942	0.910	0.926	0.997	0.992
Fuzzy string matching	0.883	1.000	0.938	0.992	0.992
Transformer-based token classification (HuBERT)	0.977	0.399	0.567	0.999	0.965

Fuzzy string matching achieved perfect recall (1.000) and the highest F1-score (0.938), reflecting its ability to capture a broad range of lexical variants in Hungarian clinical text. Regular expression–based matching demonstrated a strong balance between precision (0.942) and recall (0.910), resulting in a high F1-score (0.926). The transformer-based token classification baseline exhibited very high precision (0.977) but substantially lower recall (0.399). Simple keyword search showed limited recall (0.189), confirming its restricted sensitivity to real-world spelling and morphological variation.

3.3. Computational Performance

In terms of computational efficiency, simple keyword search was the fastest method, requiring 0.06–0.07 s per annual dataset. Regular expression–based matching required

1.62–1.64 s, while fuzzy string matching required 5.13–5.38 s. The transformer-based token classification baseline was considerably more computationally demanding, with processing times of approximately 30–35 s per annual dataset, largely due to the cost of contextual embedding generation.

Taken together, these results illustrate the trade-off between detection performance and computational cost. Fuzzy string matching achieved the highest validated recall and F1-score while maintaining execution times compatible with mid-sized clinical datasets.

The relative document-level detection rates across methods and years are illustrated in Figure 1. These results are descriptive and are not based on the manually annotated reference set.

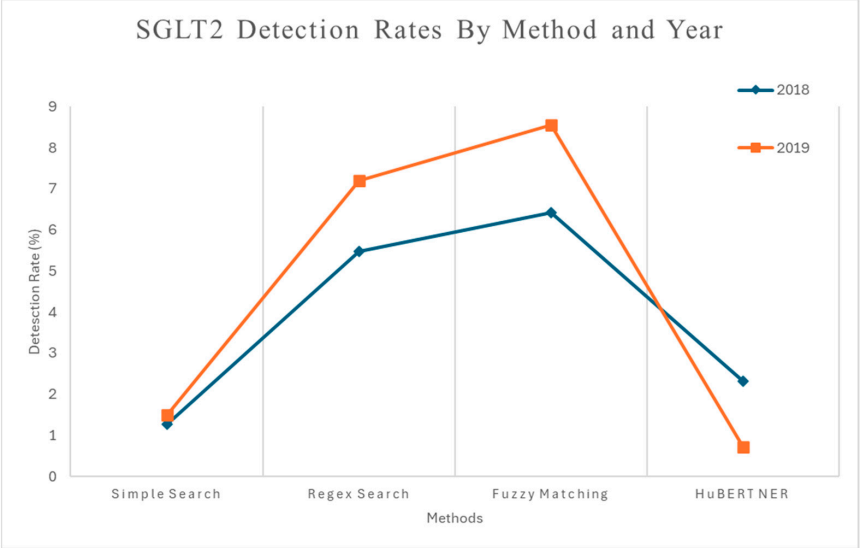

Figure 1. Document-level detection rates of SGLT2 inhibitor mentions across text-mining methods in 2018 and 2019.

4. Discussion

The results of this study are consistent with trends reported in previous work on multilingual clinical natural language processing. As observed in earlier studies, string-matching approaches do not adequately account for the linguistic variability present in human-generated clinical narratives [18]. In Hungarian clinical records, extensive morphological inflection generates multiple surface forms of medication names, which exposes a systematic limitation of literal text-matching methods. In the present analysis, simple keyword search failed to detect a substantial proportion of SGLT2 inhibitor mentions present in free-text documents. Exact matching approaches are therefore poorly suited for languages characterized by rich morphology, widespread use of abbreviations, and a lack of standardized documentation practices.

Regular expression–based matching provided a clear improvement over simple keyword search, but it also exhibited inherent limitations. Previous studies have reported that the effectiveness of regex-based medication extraction depends on coverage of anticipated variation patterns, while rare or unforeseen variants remain undetected [16]. In this study, regular expressions were designed to capture Hungarian case endings and common orthographic variants, yet they remained unable to identify uncommon misspellings and brand name variants derived from English spelling conventions, as also reported in prior work [28]. In addition, maintaining regex libraries requires continuous expert input and iterative refinement, which may limit their scalability in rapidly evolving therapeutic domains.

Fuzzy string matching consistently achieved high recall with relatively limited tuning effort. This finding aligns with international evaluations demonstrating the robustness of similarity-based methods in settings characterized by substantial orthographic variability [29]. The agglutinative nature of Hungarian further amplifies these challenges, making fuzzy matching particularly well-suited for this language. The method reliably captured inflected forms such as *dapagliflozinnal* and *empagliflozinra*, as well as non-canonical spellings. Performance remained stable across both study years, indicating resilience to variation in documentation practices [21]. These properties support the use of fuzzy string matching as a practical first-line approach for medication extraction in noisy clinical text.

The transformer-based token classification baseline demonstrated a contrasting performance profile, characterized by high precision and substantially lower recall. This pattern is consistent with reports on domain-mismatched transformer models applied to clinical text, where subword tokenization and limited domain-specific pretraining restrict recognition of specialized vocabulary [26,28]. In the present dataset, medication mentions frequently occurred in forms absent from the model's pretraining data. Shortened or colloquial variants, such as *flozin*, and heavily inflected forms were particularly prone to fragmentation into multiple subword units, reducing the availability of contextual cues required for accurate identification.

More broadly, recent deep learning research in clinical and biomedical natural language processing has emphasized the importance of contextual representation and sequence modeling for robust extraction from unstructured text [30,31]. However, such advances cannot be transferred directly to clinical text processing without substantial domain adaptation. The present findings therefore highlight a methodological trade-off. While transformer-based models offer advantages in contextual representation, they typically require extensive in-domain pretraining to achieve adequate recall in medication extraction tasks. In the absence of such adaptation, rule-based and similarity-based approaches may outperform deep learning models in real-world clinical text.

From an operational perspective, computational efficiency remains a critical consideration when processing large EHR repositories. In this study, simple keyword search, regular expression matching, and fuzzy string matching completed processing in under six seconds per annual dataset, whereas the transformer-based baseline required approximately 30 to 35 s. Although this overhead is manageable at moderate scale, it becomes increasingly relevant when applied to very large datasets or near-real-time analytical settings. Without domain-specific fine-tuning, transformer-based approaches may therefore offer a less favorable cost–benefit balance compared with lightweight rule-based alternatives.

These performance differences are particularly relevant for clinical analytics and pharmacoepidemiologic research. Misclassification of medication exposure can lead to underestimation of prevalence and biased effect estimates, potentially compromising evaluations of treatment patterns or guideline adherence. Variation in documentation practices across clinicians and departments may further exacerbate such bias if not adequately addressed during extraction. In this context, high-recall approaches are essential. Fuzzy string matching demonstrated a favorable balance between sensitivity and specificity, supporting its use as a default strategy for medication detection in morphologically rich languages.

Although hybrid extraction pipelines were not implemented in the present study, the results suggest that such approaches warrant further investigation. Two-stage architectures combining high-recall candidate generation using fuzzy or regex-based methods with contextual filtering using transformer models may provide improved precision without sacrificing sensitivity. Prior studies have shown that hybrid systems can outperform single-method pipelines in medication extraction tasks [29,32,33]. Applying such architectures to Hungarian clinical corpora, potentially supplemented with domain-specific pretrain-

ing or specialized vocabularies, may further enhance performance while maintaining computational feasibility.

Finally, the consistency of detection patterns observed across 2018 and 2019 indicates that the relative performance of the evaluated methods is robust to routine temporal variation in documentation practices. This stability supports the suitability of fuzzy string matching for practical clinical analytics and highlights the value of interpretable, low-burden NLP methods for extracting actionable information from unstructured healthcare data. As free-text documentation in health systems continues to grow, scalable and transparent approaches such as those evaluated here will remain essential for data-driven clinical research and operational analytics.

5. Limitations

This study has several limitations that should be considered when interpreting the findings. First, the data were derived from a single tertiary care institution in Hungary. Documentation practices, clinical routines, and language use may differ across hospitals and medical specialties, which may limit the generalizability of the results. External validation using datasets from multiple institutions would be required to assess the broader applicability of the observed performance patterns.

Second, manual annotation was performed for a single year (2018) and used as the reference set for calculating precision, recall, and F1-scores. While this approach is suitable for comparative method evaluation, the absence of multi-year or multi-annotator reference data limits the depth of error analysis and prevents formal assessment of inter-annotator agreement. Validation strategies could be strengthened in future work by incorporating expert annotations across multiple years and institutions. In addition, candidate documents for manual annotation were identified using keyword-assisted screening. As a result, extremely rare or unconventional mentions may not have been captured, and recall estimates, particularly for similarity-based methods, should be interpreted in the context of the reference set construction.

Third, the transformer-based token classification approach using the HuBERT backbone was evaluated without domain-specific clinical fine-tuning. Transformer models typically require large, annotated in-domain corpora to achieve optimal recall in medication extraction tasks. The lack of such domain adaptation likely contributed to the lower sensitivity observed for inflected, abbreviated, and colloquial drug mentions in this study.

Fourth, the current analysis did not explicitly address semantic phenomena such as negation, temporal qualifiers (e.g., “previously used” or “currently not taking”), or hypothetical statements. Incorporating these contextual elements would require additional rule-based logic or more advanced contextual modeling beyond the scope of the present comparison.

Finally, the analysis focused exclusively on SGLT2 inhibitors as a single drug class. While this allowed detailed methodological evaluation, extrapolation of the findings to other medication classes, diagnostic terms, or procedural terminology would require additional comparative analyses.

6. Conclusions

Text-mining methods were evaluated and compared in this study for the identification of SGLT2 inhibitor mentions in unstructured Hungarian clinical text. Among the evaluated approaches, fuzzy string matching demonstrated the most favorable trade-off between sensitivity, computational efficiency, and practical applicability. Regular expression–based matching also achieved strong performance, albeit with a moderate configuration and

maintenance effort. Transformer-based token classification showed high precision but low sensitivity when applied without domain-specific fine-tuning.

These findings underscore the continued relevance of lightweight and interpretable text-mining methods for practical clinical analytics, particularly in morphologically rich languages such as Hungarian. The results support the use of fuzzy string matching as an efficient front-line solution for medication extraction and as a suitable foundation for future hybrid pipelines that combine high-recall rule-based candidate generation with contextual deep learning models.

Funding: This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the University of Debrecen (5610-2020, 17 December 2020).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from Clinical Centre of the University of Debrecen, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Clinical Centre of the University of Debrecen, Hungary.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.-Z. Big data for health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1193–1208. [CrossRef]
2. Baro, E.; Degoul, S.; Beuscart, R.; Chazard, E. Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Res. Int.* **2015**, *2015*, 639021. [CrossRef]
3. Chin, L.; Khozin, S. A digital highway for data fluidity and data equity in precision medicine. *Biochim. Biophys. Acta (BBA)-Rev. Cancer* **2021**, *1876*, 188575. [CrossRef]
4. Padoan, A.; Plebani, M. Flowing through laboratory clinical data: The role of artificial intelligence and big data. *Clin. Chem. Lab. Med.* **2022**, *60*, 1875–1880. [CrossRef]
5. Martin-Sanchez, F.; Verspoor, K. Big Data in Medicine is Driving Big Changes. *Yearb. Med. Inform.* **2014**, *9*, 14–20.
6. Sedda, G.; Gasparri, R.; Spaggiari, L. Challenges and Innovations in Personalized Medicine Care. *Future Oncol.* **2019**, *15*, 3305–3308. [CrossRef] [PubMed]
7. Luque, C.; Luna, J.M.; Luque, M.; Ventura, S. An advanced review on text mining in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1302. [CrossRef]
8. Gasparri, R.; Sedda, G.; Spaggiari, L. Comment from the Editor to the Special Issue: "Big Data and Precision Medicine Series I: Lung Cancer Early Diagnosis". *J. Clin. Med.* **2018**, *7*, 28. [CrossRef] [PubMed]
9. Nagy, A.; Kovács, N.; Pálinkás, A.; Sipos, V.; Vincze, F.; Szőllősi, G.; Ádány, R.; Czifra, Á.; Sándor, J. Improvement in Quality of Care for Patients with Type 2 Diabetes in Hungary Between 2008 and 2016: Results from Two Population-Based Representative Surveys. *Diabetes Ther.* **2019**, *10*, 757–763. [CrossRef]
10. ten Hoope, S.; Welvaars, K.; van Geijtenbeek, K.; Klok-Everaars, M.; van Schaik, S.; Karapinar-Çarkit, F. Applying text-mining to clinical notes: The identification of patient characteristics from electronic health records (EHRs). *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 302. [CrossRef] [PubMed]
11. Karystianis, G.; Sheppard, T.; Dixon, W.G.; Nenadic, G. Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC Med. Inform. Decis. Mak.* **2016**, *16*, 18. [CrossRef]
12. Extraction of Medication and Temporal Relation from Clinical Text Using Neural Language Models | IEEE Conference Publication | IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/10386489 (accessed on 28 November 2025).
13. Kim, Y.; Meystre, S.M. Ensemble method–based extraction of medication and related information from clinical texts. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 31–38. [CrossRef] [PubMed]

14. Bailey, C.J.; Day, C.; Bellary, S. Renal Protection with SGLT2 Inhibitors: Effects in Acute and Chronic Kidney Disease. *Curr. Diabetes Rep.* **2022**, *22*, 39–52. [CrossRef] [PubMed]
15. Cowie, M.R.; Fisher, M. SGLT2 inhibitors: Mechanisms of cardiovascular benefit beyond glycaemic control. *Nat. Rev. Cardiol.* **2020**, *17*, 761–772. [CrossRef]
16. Karagiannis, T.; Tsapas, A.; Athanasiadou, E.; Avgerinos, I.; Liakos, A.; Matthews, D.R.; Bekiari, E. GLP-1 receptor agonists and SGLT2 inhibitors for older people with type 2 diabetes: A systematic review and meta-analysis. *Diabetes Res. Clin. Pract.* **2021**, *174*, 108737. [CrossRef]
17. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
18. Siklósi, B. Methods for Processing Noisy Texts and their Application to Hungarian Clinical Notes. Ph.D. Thesis, Pázmány Péter Catholic University, Budapest, Hungary, 2015.
19. Mahtta, D.; Ramsey, D.J.; Lee, M.T.; Chen, L.; Al Rifai, M.; Akeroyd, J.M.; Vaughan, E.M.; Matheny, M.E.; Santo, K.R.d.E.; Navaneethan, S.D.; et al. Utilization Rates of SGLT2 Inhibitors and GLP-1 Receptor Agonists and Their Facility-Level Variation Among Patients With Atherosclerotic Cardiovascular Disease and Type 2 Diabetes: Insights From the Department of Veterans Affairs. *Diabetes Care* **2022**, *45*, 372–380. [CrossRef]
20. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. ScienceDirect. Available online: https://www.sciencedirect.com/book/edited-volume/9780123869791/practical-text-mining-and-statistical-analysis-for-non-structured-text-data-applications (accessed on 28 November 2025).
21. Jouffroy, J.; Feldman, S.F.; Lerner, I.; Rance, B.; Burgun, A.; Neuraz, A. Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study. *JMIR Med. Inform.* **2021**, *9*, e17934. [CrossRef]
22. 3.12.9 Documentation. Available online: https://docs.python.org/3.12/ (accessed on 2 March 2025).
23. Leaman, R.; Khare, R.; Lu, Z. Challenges in Clinical Natural Language Processing for Automated Disorder Normalization. *J. Biomed. Inform.* **2015**, *57*, 28–37. [CrossRef] [PubMed]
24. Sohn, S.; Clark, C.; Halgrim, S.R.; Murphy, S.P.; Chute, C.G.; Liu, H. MedXN: An open source medication extraction and normalization tool for clinical text. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 858–865. [CrossRef]
25. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
26. SZTAKI HLT | Natural Language Processing Methods for Language Modeling. Available online: https://hlt.bme.hu/en/publ/nemeskey_2020 (accessed on 10 October 2025).
27. *Stata Statistical Software*, version 18; StataCorp LLC: College Station, TX, USA, 2023.
28. Perera, N.; Dehmer, M.; Emmert-Streib, F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Front. Cell Dev. Biol.* **2020**, *8*, 673. [CrossRef] [PubMed]
29. Bhasuran, B.; Murugesan, G.; Abdulkadhar, S.; Natarajan, J. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. *J. Biomed. Inform.* **2016**, *64*, 1–9. [CrossRef] [PubMed]
30. Zhang, Y.; Xiao, Y.; Zhang, Y.; Zhang, T. Video saliency prediction via single feature enhancement and temporal recurrence. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111840. [CrossRef]
31. Zhang, Y.; Wang, T.; Xue, L.; Lian, W.; Tao, R. ORSI Salient Object Detection via Pro-gressive Interaction and Saliency-Guided Enhancement. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 6002105. [CrossRef]
32. Chen, A.; Yu, Z.; Yang, X.; Guo, Y.; Bian, J.; Wu, Y. Contextualized medication information extraction using Transformer-based deep learning architectures. *J. Biomed. Inform.* **2023**, *142*, 104370. [CrossRef]
33. Campillos-Llanos, L.; Valverde-Mateos, A.; Capllonch-Carrión, A. Hybrid natural language processing tool for semantic annotation of medical texts in Spanish. *BMC Bioinform.* **2025**, *26*, 7. [CrossRef]