Mapping Medical Procedure Codes Using Language Models

Mariša Ratajec
University of Ljubljana, Faculty of Electrical Engineering; Jožef Stefan Institute
Ljubljana, Slovenia
ratajec.marisa@gmail.com

Anton Gradišek
Jožef Stefan Institute
Ljubljana, Slovenia
anton.gradisek@ijs.si

Nina Reščič*
Jožef Stefan Institute
Ljubljana, Slovenia
nina.rescic@ijs.si

Abstract

Aligning medical procedure codes across national classification systems is a challenging task. We investigate the mapping of Slovenian KTDP expressions to German OPS codes using fuzzy matching, biomedical language models (BioBERT, GatorTron), a hybrid approach, and ChatGPT. In the absence of ground truth, we assess consistency across methods and conduct manual reviews. Results show that differences in code structure and expression detail pose major barriers to alignment. Expert validation will be essential for improving accuracy.

Keywords

procedure coding, KTDP, OPS, semantic similarity, BioBERT, fuzzy matching, GatorTron, ChatGPT

1 Introduction

Different countries maintain their own national classification systems for medical procedures, used for clinical documentation, reimbursement, public reporting, and statistical analysis. In Slovenia, healthcare professionals rely on a domestic procedural coding system, while in Germany, the Operationen- und Prozedurenschlüssel (OPS) is used.

At the University Medical Centre (UMC) Ljubljana in Slovenia, interest has emerged in matching expressions from the Klasifikacija terapevtskih in diagnostičnih postopkov in posegov (KTDP) with the German OPS classification system. The purpose is to allow international reporting, cost estimation, and comparative analysis of healthcare procedures.

1.1 Problem Outline

Aligning Slovenian procedural expressions with German OPS codes is a complex task. The Slovenian dataset contains approximately 6,000 expressions, whereas the German OPS classification includes more than 60,000 distinct entries, covering multiple levels of specificity in various medical domains. Manual mapping is time-consuming and impractical, primarily due to the size of datasets and the absence of convenient tools for efficient code retrieval and comparison.

To address this challenge, we explored the development of computational approaches to support and accelerate the mapping process. Due to the nature of the data and the semantic variation between codes, we tested several techniques, including fuzzy string matching, semantic similarity scoring, and large language models (LLMs), such as BioBERT, GatorTron, and OpenAI models. We also explored a hybrid approach that integrates fuzzy matching with LLM-derived semantic embeddings.

In this paper, we present the application of the selected methods for aligning Slovenian KTDP procedure expressions with German OPS codes. We evaluate their performance, limitations and discuss key challenges associated with this type of code matching problem.

2 Methodology

2.1 Datasets

2.1.1 Slovenian Dataset. The Slovenian dataset is based on the Klasifikacija terapevtskih in diagnostičnih postopkov in posegov (KTDP)[6], version 11, which has been officially implemented nationwide since 1 January 2023. This classification system is used to code medical procedures in all levels of healthcare in Slovenia and is structurally aligned with the Australian Classification of Health Interventions (ACHI), adapted to the local context.

KTDP consists of 20 chapters, each covering a different clinical domain. The chapters are organised primarily by body system (e.g. nervous, endocrine, cardiovascular), with additional sections dedicated to dental care, imaging services, radiation oncology, and interventions not elsewhere classified. Each chapter is subdivided into multiple blocks, which group related procedures under shared headings.

In total, the classification includes approximately 6,000 unique procedures. Each is assigned a specific code in a structured numeric format composed of a five-digit base and a two-digit extension (e.g. 36564-00).

2.1.2 German Dataset. The German dataset is based on Operationen und Prozedurenschlüssel (OPS), version 2024 [2], which is officially used nationwide for coding medical procedures. Maintained by the Federal Institute for Drugs and Medical Devices (BfArM), OPS is revised annually. It is derived from the WHO's International Classification of Procedures in Medicine (ICPM) and adapted to the German healthcare system.

The classification is organised into six main chapters, covering the following clinical domains: diagnostic measures (Chapter 1), imaging diagnostics (Chapter 3), surgical procedures (Chapter 5), medications (Chapter 6), non-operative therapeutic measures (Chapter 8), and supplementary measures (Chapter 9). Each chapter is further subdivided into categories and blocks, which group related procedures based on functional or anatomical criteria.

OPS comprises approximately 60,000 unique procedures. Each is assigned a hierarchical alphanumeric code, consisting of a four-digit base and optional numeric or alphanumeric extensions (e.g. 5-384.50 or 8-844.5c). The coding system follows a structured hierarchy, beginning with the chapter number (e.g. 5 for surgical procedures), followed by a category (e.g. 5-38 for vascular

[*] Corresponding author

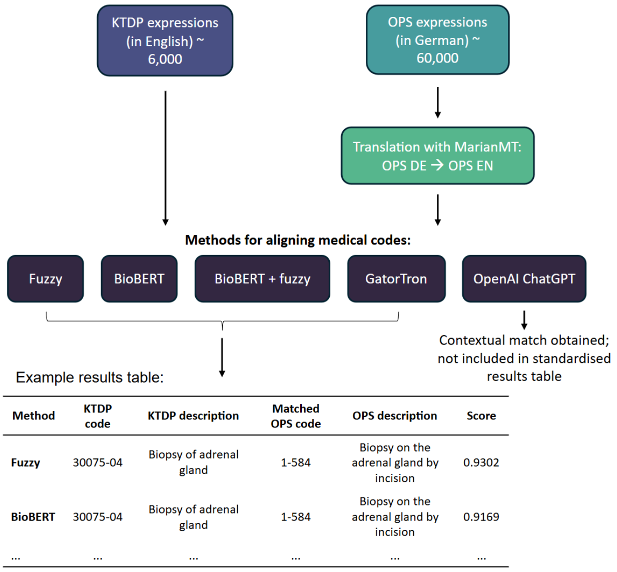
Information Society 2025, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2025.skui.9648

surgery) and subcategories (e.g. 5-384 for specific surgical techniques). The digits and characters after the dot denote the exact intervention.

2.1.3 Differences and Similarities between Datasets. Although both classification systems serve a similar purpose, they differ in structure and level of detail. The German dataset includes very specific and thoroughly described procedures, clearly outlining each individual service. The Slovenian system, on the other hand, uses broader and more general descriptions, without the same amount of detail or length.

Moreover, there is limited direct lexical overlap between the two datasets. Even when procedures are conceptually similar, their descriptions often differ in phrasing, level of specificity, or use of synonyms. As a result, one-to-one matching is rarely straightforward and requires both structural alignment and semantic interpretation.

2.2 Pipeline


KTDP expressions (in English) ~ 6,000
OPS expressions (in German) ~ 60,000
Translation with MarianMT: OPS DE → OPS EN
Methods for aligning medical codes:
Fuzzy
BioBERT
BioBERT + fuzzy
GatorTron
OpenAI ChatGPT
Contextual match obtained; not included in standardised results table
Example results table:

Method	KTDP code	KTDP description	Matched OPS code	OPS description	Score
Fuzzy	30075-04	Biopsy of adrenal gland	1-584	Biopsy on the adrenal gland by incision	0.9302
BioBERT	30075-04	Biopsy of adrenal gland	1-584	Biopsy on the adrenal gland by incision	0.9169
...	...	...	...	...	...

Figure 1: Overview of the matching pipeline and example results. KTDP expressions in English were aligned to translated OPS expressions using five methods: fuzzy matching, BioBERT, a combined BioBERT+fuzzy approach, GatorTron, and OpenAI ChatGPT. All methods except ChatGPT produced structured outputs with match scores, as shown in the example results table. ChatGPT returned only contextual matches without comparable scoring and was therefore excluded from the standardised evaluation table.

The overall process is summarised in a pipeline diagram (Figure 1), which outlines each step — from dataset preparation and translation to the application of matching methods and the structure of resulting outputs. Each component of the pipeline is described in detail in the following subsections.

2.2.1 Translation. Since Slovenian KTDP expressions were already available in English, the German OPS procedure names were translated to English to enable semantic comparison. For this purpose, we used the MarianMT model (`Helsinki-NLP/opus-mt-de-en`) [4], a transformer-based neural machine translation model. Although not specifically fine-tuned for clinical texts, MarianMT has demonstrated strong performance in medical translation tasks, particularly for structured terminology [5], making it a suitable and practical choice for this application.

2.2.2 Language-based code pairing. To perform code matching, we initially applied a language-based code pairing approach using fuzzy matching, implemented via the RapidFuzz library [1]. Fuzzy matching is particularly useful in cases where expressions differ slightly in wording, structure, or spelling. We applied the token set ratio, which compares the sets of unique words in two strings and calculates a similarity score based on the overlap of unique tokens, with edit distance applied to the remaining unmatched parts. This method is insensitive to word order and robust to minor variations in phrasing. Using this approach, each English KTDP expression was compared with all translated OPS descriptions. For each KTDP entry, we selected the best matching OPS procedure based on the highest fuzzy similarity score and recorded the corresponding code, description, and score for further analysis.

2.2.3 Semantic-based code pairing. As a second approach, we applied a semantic-based code pairing approach using contextual embeddings derived from transformer-based language models. Specifically, we tested two pretrained models: `pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb` [3], a SentenceTransformer variant of BioBERT fine-tuned on biomedical and inference tasks, and `UFNLP/gatortron-base` [10], a GatorTron model pre-trained on large-scale clinical corpora. Both models were selected for their strong performance in biomedical language understanding [7] and to investigate how model choice influences the quality of semantic code alignment.

Using each model, both KTDP expressions and translated OPS descriptions were encoded into dense semantic vectors. Cosine similarity was then computed between each KTDP embedding and all OPS embeddings to assess semantic closeness. As in the previous approach, the top matching OPS procedure for each KTDP expression was selected and recorded following the same procedure as before.

2.2.4 Combined code pairing. In addition to the individual use of semantic and lexical methods, we implemented a hybrid matching approach that combines the strengths of both. Specifically, we integrated semantic similarity scores obtained from BioBERT embeddings with lexical similarity scores derived from fuzzy matching (token set ratio). For each KTDP expression, both similarity measures were computed independently against all translated OPS descriptions. The final similarity score for each pair was calculated as a weighted average:

$$\text{score}_{\text{final}} = w_{\text{semantic}} \cdot \text{score}_{\text{semantic}} + w_{\text{lexical}} \cdot \text{score}_{\text{lexical}}$$

We experimented with two weighting schemes: one with equal weights ($w_{\text{semantic}} = 0.5$, $w_{\text{lexical}} = 0.5$) and another prioritising semantic similarity ($w_{\text{semantic}} = 0.7$, $w_{\text{lexical}} = 0.3$), to assess how different balances influence match quality. For each KTDP expression, the OPS description with the highest combined score was selected and recorded along with the corresponding code and similarity score.

This approach was motivated by practical observations in the literature, where combining surface-level and context-aware similarity often yields more robust results, especially in cases where purely semantic models overlook minor wording differences or where lexical methods fail to capture deeper conceptual alignment [9].

2.2.5 ChatGPT code pairing. As a final exploratory method, we used a custom ChatGPT instance (GPT-4o, OpenAI) [8] to evaluate the potential of conversational large language models (LLMs) for code matching. We uploaded all relevant documentation, including KTDP expressions, translated OPS procedures, and background materials, to a private GPT environment. For each KTDP entry, we either asked the model to suggest the best-matching OPS procedure directly or first requested an interpretation of the KTDP term followed by a context-based match. This approach allowed us to assess whether ChatGPT's contextual reasoning could complement or outperform traditional embedding-based or lexical matching methods.

3 Evaluation

The absence of a validated ground truth presents a fundamental challenge in assessing the quality of our matching results. Without expert clinical validation, it is unclear how accurate individual matches are or which method performs best. To address this, we first conducted a broad quantitative analysis to evaluate consistency, disagreement, and similarity across methods. These metrics provide indirect but informative insights into model behaviour, helping to characterise matching patterns even in the absence of formal validation. Following this initial assessment, we performed a small-scale manual review to better understand the plausibility of selected matches. We examined examples with both high and low matching scores, identifying cases of clear agreement as well as notable mismatches. This informal inspection offered additional intuition on method performance and highlighted the need for domain expertise to reliably judge alignment quality.

To begin the quantitative evaluation, we examined how often different methods assigned KTDP expressions to the same general procedural category. To do this, we compared the prefixes of the top-1 matched OPS codes across all methods, where the prefix corresponds to the first digit of the OPS code and indicates the high-level category of the procedure (e.g., diagnostic, surgical, therapeutic). This allowed us to assess agreement at a broader level, independent of specific code details.

The results revealed a relatively high degree of consistency: in 64.2% of cases (n = 4000), all methods returned OPS codes with the same prefix, indicating agreement on the general procedural category. In the remaining 35.8% of cases (n = 2231), there was partial agreement - some methods aligned on the prefix, while others diverged. Notably, there were no cases in which all methods assigned entirely different prefixes, suggesting that at least a minimal level of agreement was always preserved at the category level.

However, when comparing full OPS codes, agreement dropped substantially. Only 2.9% of cases (n = 178) exhibited full consensus across all methods. Most cases (90.1%, n = 5613) fell into the "some same" category, where at least two methods agreed, and 7.1% (n = 440) showed complete disagreement, with each method proposing a different code. These results indicate that, while methods often converge on the general category of a procedure, they frequently differ in the specific code they assign within that category.

To further examine how the methods differ in their assignment behaviour, we analysed the distribution of top-1 matched OPS codes across the six main procedural chapters. As illustrated in Figure 2, all methods predominantly mapped KTDP expressions to Chapter 5 (surgical procedures), reflecting the procedural nature of the source data. In contrast, assignments to Chapter 6 (medications) and Chapter 9 (supplementary measures) were relatively infrequent. This general distribution pattern was consistent across methods, indicating a shared tendency to favour procedural codes.

Even so, some notable differences were observed. For example, GatorTron assigned fewer expressions to Chapter 5 compared to the other methods and exhibited a relatively higher proportion of matches to Chapter 8 (non-operative therapeutic measures). Manual review of these cases revealed that many of the expressions lacked a clearly corresponding OPS code, which may have led the model to prefer broader categories. Still, in the absence of expert validation, we cannot determine whether such assignments are more or less accurate.

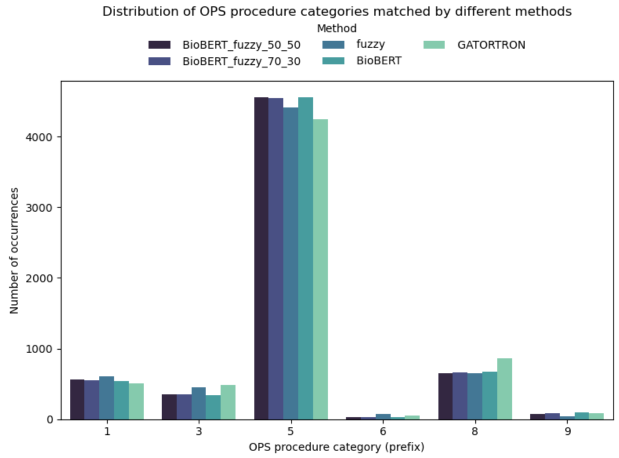


Figure 2: Distribution of top-1 matched OPS codes across the six main procedural chapters for each matching method. Chapter 1 represents diagnostic measures, Chapter 3 imaging diagnostics, Chapter 5 surgical procedures, Chapter 6 medications, Chapter 8 non-operative therapeutic measures, and Chapter 9 supplementary measures.

To investigate whether certain KTDP procedures are inherently easier to match due to wording or alignment with OPS terminology, we analysed the standardised match score values across all methods using a heatmap (Figure 3). The goal was to determine whether consistent scoring patterns could help identify procedures that are generally easier or more difficult to match, regardless of the specific method used.

The heatmap displays Z-standardised scores for each method, with expressions sorted by BioBERT scores. Although we expected some consistency (i.e., easier expressions receiving higher scores across all methods and harder ones receiving lower scores), the results showed considerable variation. In many cases, a procedure scored higher with one method and lower with another, suggesting that matching difficulty is method-dependent and influenced by how each approach interprets textual or structural similarity.

Notably, BioBERT and the hybrid BioBERT-fuzzy method produced very similar score profiles. GatorTron and fuzzy approach showed more divergence, indicating different sensitivities to terminology structure, dataset alignment, or surface-level phrasing.

This suggests that methods differ not only in which codes they select, but also in how confidently they make those matches.

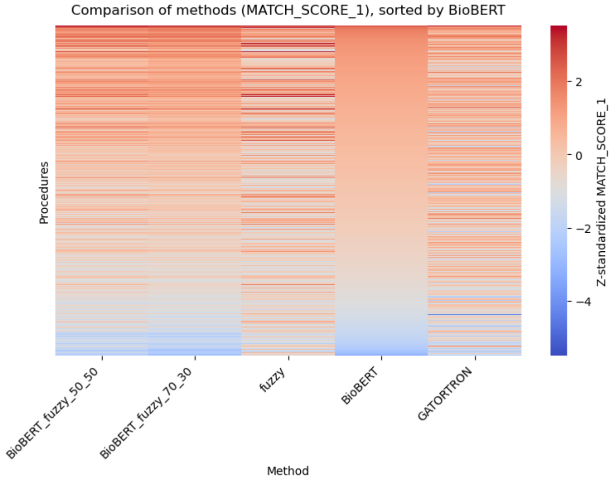


Figure 3: Heatmap of Z-standardised MATCH_SCORE_1 values across all KTDP expressions, sorted by BioBERT scores. The plot illustrates variation in score strength across methods, highlighting differences in confidence and matching behaviour.

After developing a broader understanding of inter-method differences through quantitative analyses, we conducted a focused manual review of selected examples to qualitatively assess the plausibility of top matches. We examined expressions with both high and low matching scores across methods to explore whether any consistent patterns could be observed.

For expressions with high scores and full agreement across methods, the matches were typically straightforward: the KTDP expression was either identical or highly similar to an OPS entry, often requiring no complex interpretation. These cases tended to represent procedural descriptions that appeared in both datasets with minimal variation.

In contrast, lower-scoring expressions revealed more complex challenges. Two main issues emerged during manual inspection. First, several KTDP procedures had no direct equivalent in the OPS system because they are typically recorded in other coding systems (e.g., vaccinations or disease-specific protocols). Second, many KTDP expressions were written in a general or aggregated form, often combining multiple procedural steps into a single description. OPS, on the other hand, is highly granular, with detailed and precisely defined codes. As a result, some KTDP expressions may correspond to multiple distinct OPS codes, or only partially align with available entries.

These observations suggest that performance limitations are not solely attributable to matching algorithms themselves, but also to structural mismatches and representational differences between the source datasets. This highlights a key challenge in aligning procedural coding systems across countries.

3.1 ChatGPT

Despite leveraging ChatGPT's capacity for contextual reasoning by first interpreting the KTDP expression and then performing the match, the resulting OPS codes were, in most cases, identical to those produced by previously described methods. This suggests that the added interpretation step did not substantially improve matching performance. As previously discussed, this outcome likely reflects the inherent differences in datasets.

4 Conclusion

Our study highlights the considerable challenge of aligning procedural coding systems across countries with different documentation practices. Despite employing a range of computational methods (ranging from fuzzy matching and semantic embeddings to large language models) the observed differences in dataset structure and content significantly limited matching performance. In particular, the lack of detail in some KTDP expressions, the high specificity of OPS codes, and the absence of one-to-one equivalents all contributed to inconsistent or ambiguous results.

Crucially, no ground truth currently exists to objectively evaluate the quality of these matches. Although indirect metrics and manual inspection provide useful information, they cannot replace expert validation. Therefore, the most important next step is to involve medical professionals in generating a gold standard reference set. This would enable formal benchmarking of different methods and support the development of more reliable and generalisable code alignment pipelines in the future.

Ultimately, our findings suggest that the key limitation lies not in the technical capability of the methods themselves, but in the fundamental heterogeneity of the datasets and the differing philosophies of procedural encoding. Addressing this mismatch will be essential for any future efforts to enable international interoperability of procedural coding systems.

Acknowledgments

The authors are grateful for the valuable input and ideas contributed by the medical team of the University Medical Center Ljubljana.

Funding

This work was supported by the Slovenian Research and Innovation Agency (Research Core Funding Number P2-0209).

References

[1] [SW] Max Bachmann, rapidfuzz/RapidFuzz: Release 3.13.0 version v3.13.0, Apr. 2025. doi: 10.5281/zenodo.15133267, url: https://doi.org/10.5281/zenodo.15133267.

[2] Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). 2023. *Operationen- und Prozedurenschlüssel (OPS), Version 2024: Internationale Klassifikation der Prozeduren in der Medizin – Systematisches Verzeichnis*. BfArM. Bonn, Germany.

[3] Pritam Deka, Anna Jurek-Loughrey, et al. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*. Springer, 3–15.

[4] Marcin Junczys-Dowmunt et al. 2018. Marian: Fast Neural Machine Translation in C++. Tech. rep. arXiv:1804.00344. Demonstration paper, version v3. arXiv, (Apr. 2018). doi: 10.48550/arXiv.1804.00344.

[5] Bunyamin Keles, Murat Gunay, and Serdar Caglar. 2024. LLMs-in-the-loop Part-1: Expert Small AI Models for Bio-Medical Text Translation. Tech. rep. arXiv:2407.12126. Preprint. arXiv, (July 2024). doi: 10.48550/arXiv.2407.12126.

[6] Nacionalni inštitut za javno zdravje (NIJZ). 2023. *Klasifikacija terapevtskih in diagnostičnih postopkov in posegov: Pregledni seznam (Verzija 11)*. NIJZ. Ljubljana, Slovenia.

[7] Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: a review. *Informatics*, 11, 3, 57. doi: 10.3390/informatics11030057.

[8] OpenAI. 2024. Gpt-4o. Accessed: August 2025. (2024). https://openai.com/index/gpt-4o.

[9] Mohammed Suleiman Mohammed Rudwan and Jean Vincent Fonou-Dombeu. 2023. Hybridizing fuzzy string matching and machine learning for improved ontology alignment. *Future Internet*, 15, 7, 229. doi: 10.3390/fi15070229.

[10] Xi Yang et al. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5, 1, 194.