Extracting Medication Instructions from Dutch General Practice Electronic Health Records with Local Natural Language Processing

Marya Dukmak, Constanza L Andaur Navarro, Artuur M Leeuwenberg
University Medical Center Utrecht, Utrecht University,
Heidelberglaan 100, 3508 GA Utrecht, The Netherlands
maryadukmak@gmail.com, candaurn@umcutrecht.nl, aleeuw15@umcutrecht.nl

Abstract

The extraction of structured medication prescription data from unstructured clinical text remains a critical challenge for clinical research and data standardization. This study investigates the application of Natural Language Processing (NLP) techniques to Dutch electronic health records (EHRs) from the Julius General Practitioners Network. The goal is to automatically extract key prescription attributes including dosage, duration, and medication unit and prepare them for integration into the ConcePTION Common Data Model, to support scalable pharmacoepidemiological research. We compare a lightweight rule-based system with transformer-based models (RobBERT and MedRoBERTa) under the technical constraints of a Trusted Research Environment, where external resources and cloud-based solutions are restricted. Using a dataset of 1,819 manually annotated records, the approaches are evaluated on predictive performance and computational costs. Results show that the rule-based system achieves strong accuracy and computational costs for structured patterns, while transformer-based models demonstrate greater robustness to linguistic variability. However, both approaches encounter difficulties with ambiguous dosage formats and long treatment durations. Our findings indicate that NLP methods can substantially improve the structuring of Dutch prescription data and support scalable pharmacoepidemiological research. Future work should focus on improving generalization and expanding annotated datasets to enhance model reliability.

Keywords: Dutch Electronic Health Records (EHRs), NLP, Medication information extraction

1. Introduction

Electronic Health Records (EHRs) have become a fundamental component of modern healthcare, enabling the large-scale collection of patient data and supporting clinical decision-making as well as medical research. However, a substantial portion of EHR data remains unstructured in free text form, limiting its usability for secondary use. An important secondary use of EHR data is the conduct of pharmacoepidemiological studies that investigate medication safety after their introduction into healthcare practice. Critically, these large-scale (often international) studies rely on common data models (CDMs): data from all participating healthcare centers - with often different local recording practices - are first mapped to the same data structure and semantics, to ensure comparability and usability for research. This article investigates to what degree different natural language processing (NLP) methods can extract relevant medication information from EHR free text, to make them available for pharmacoepidemiological research. Specifically, we focus on the extraction of medication prescription instructions from Dutch general practice EHRs.

1.1. Case study setting

Having accurate data about what medication is prescribed to each patient is essential for pharmacoepidemiological studies that evaluate treatment strategies and medication safety. One of the most used common data models to encode such information used in this field is the ConcePTION CDM (Thurin et al., 2022). We focus on three essential ConcePTION CDM variables in this study: dosage, duration, and medication units. Figure 1 illustrates an example of the input and outputs for the NLP task discussed in the current study.

A second important aspect of this setting is the use of (non-English) Dutch GP EHR data from the Julius General Practitioners Network (Smeets et al., 2018), which has important implications for what NLP techniques can be used: they should handle Dutch clinical free text (with very specific terminology and abbreviations). And, importantly, the use of these (pseudonimized) sensitive data is only permitted in a strict privacy protected digital research environment (DRE): data may not be send to third parties (e.g., proprietary LLM providers) and computational resources are relatively limited. In other words, NLP techniques are constrained to locally runnable arguably smaller models that can handle Dutch clinical text: these are thus the focus of this study.

1.2. Research contributions

Within this case study, we compare the following NLP methods: a rule-based approach with regular expressions, and two Dutch transformer-based models: RobBERT and MedRoBERTa.nl. Our research questions pertain to how well each of these methods extracts the relevant prescription informa-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 174–182
16 May 2026.

tion and what is the computational costs.

This paper makes the following contributions:

- We present an NLP pipeline for extracting key prescription attributes dosage, duration, and medication unit from Dutch primary care EHR data.
- We provide a comparative evaluation of rule-based and transformer-based approaches within the constraints of a secure digital research environment.
- We analyze the strengths and limitations of both approaches in terms of performance, efficiency, and robustness to prescribing variability.

By improving the structuring of prescription data, this work supports the development of scalable and reusable clinical datasets, ultimately facilitating large-scale pharmacoepidemiological research.

2. Related Work

Medical information extraction aims to transform large volumes of unstructured clinical text into structured data suitable for analysis. As the adoption of Electronic Health Records has increased, automated extraction methods have become essential for efficiently identifying clinically relevant information (Ford et al., 2016).

Early approaches relied primarily on rule-based systems that leverage handcrafted grammars, lexicons, and domain knowledge. These methods have demonstrated strong performance in structured clinical settings. For example, (Mykowiecka et al., 2009) developed a rule-based system for extracting attributes from Polish clinical documents, achieving precision and recall above 99%. Despite their reliability and interpretability, rule-based approaches often struggle with scalability and require continuous manual updates to capture linguistic variation.

To address these limitations, machine learning techniques such as Conditional Random Fields (CRFs) and Support Vector Machines have been introduced for clinical information extraction. (Patrick and Li, 2009) presented a hybrid system combining rule-based preprocessing with statistical classifiers, achieving an F-score of 93.5% for medication extraction from emergency department notes. Similarly, systems integrating probabilistic models with linguistic rules have shown that combining domain expertise with data-driven methods can improve extraction performance.

More recently, transformer-based architectures have substantially advanced clinical NLP. Pre-trained language models enable the capture of contextual relationships within complex clinical narratives while reducing the need for manual feature engineering (Yang et al., 2020). These models have been successfully applied to organize electronic medical records and support clinical decision-making (Landolsi et al., 2022). However, challenges remain, particularly regarding domain adaptation and the handling of specialized medical terminology (Tayefi et al., 2021).

While medication information extraction methods with transformer models have recently been explored for non-English (Fabacher et al., 2025; Campillos-Llanos et al., 2025), the majority of studies and transformers primarily utilized English corpora, and these results or models cannot be directly generalized to Dutch medical text without evaluation or adaptation. Research on Dutch clinical NLP remains comparatively limited(AlShuweihi et al., 2021). Addressing this gap, (Muizelaar et al., 2024) evaluated BERT-based models for extracting lifestyle characteristics from Dutch clinical notes, demonstrating the effectiveness of domain-specific pre-training.

A key resource in this area is MedRoBERTa.nl, the first transformer-based language model trained specifically on Dutch electronic health records (Verkijk and Vossen, 2021). The model outperforms general-purpose Dutch language models on several clinical NLP tasks, highlighting the importance of in-domain language resources for under-represented languages.

Together, these studies indicate a shift from deterministic extraction toward context-aware language models, while also emphasizing the continued relevance of interpretable approaches in clinical environments. Building on this body of work, the present study compares rule-based and transformer-based strategies for extracting structured medication information from Dutch primary care EHRs.

3. Methodology

3.1. Data and annotation

The data used in this study originates from the Julius General Practitioners Network (JGPN) (Smeets et al., 2018), a large primary care database made available for medical research, covering over 390 general practitioners in the Utrecht region, spanning from years 1919 to 2024, and containing over three million medication records.

Data were first filtered for recency (after year 2000) and availability of the values for the relevant columns (prescription instruction and product strength). Then, from the remaining 861,515 records, a random subset of 2,000 prescription records was selected to represent real-world clinical

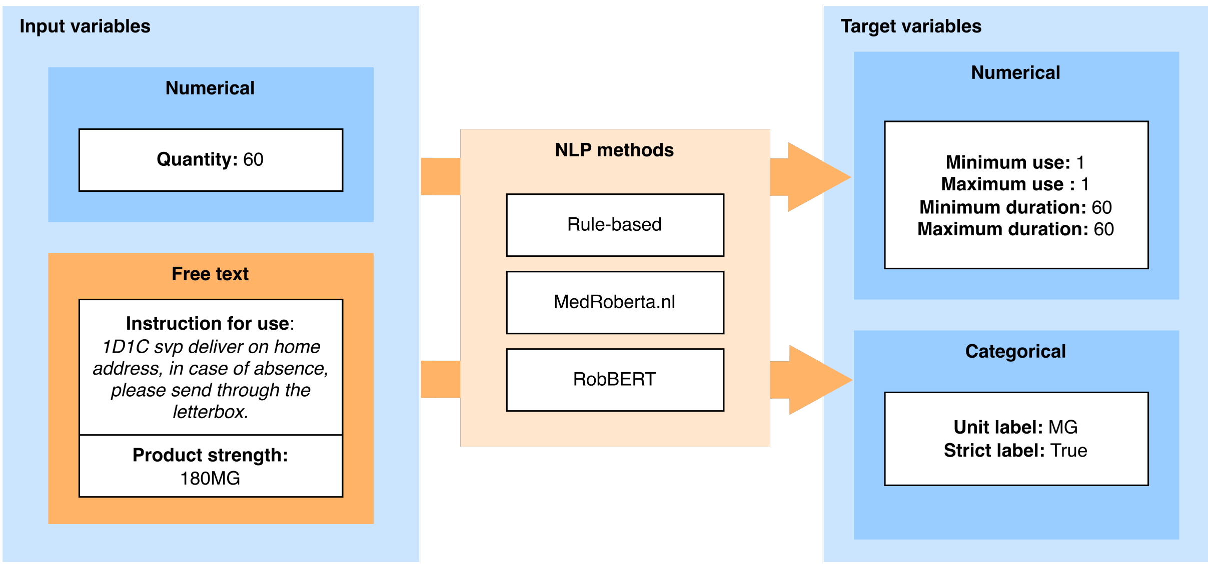

Figure 1: Schematic input output overview for an (artificial, English-language) example.

documentation practices. The subset includes both frequently occurring and rare dosage patterns, and no data balancing was applied in order to preserve the natural distribution of prescription instructions.

Using LabelStudio (Tkachenko et al., 2020–2025) each record was first annotated by a single annotator (Dutch native, non-clinical, data scientist) for key medication attributes: dosage, duration, and medication unit (descriptions found in Table 1. Any uncertain cases were held out and checked with a pharmacist or general practitioner, depending on the source of uncertainty. A random subset of 100 records was annotated by a second annotator (Dutch native, non-clinical, data scientist) to calculate inter-annotator agreement, which can be found in Table 2. A screenshot of the annotation tool and the LabelStudio configuration are added to the supplementary materials.

Finally, after excluding cases the pharmacist and GP could not label (e.g., sometimes interpretation required the full patient journey instead of only the prescription instruction), 1,819 were selected to develop and evaluate the NLP models.

3.2. Experimental setup

The 1,819 annotated records were split in a train (80%) and test set (20%) at the GP practice level (to prevent leakage from having the same GP, with likely more similar documentation style in the train and test set). For each of the three modeling methods, the train set was used for its development (also for the rule-based system) and test set for its evaluation.

3.2.1. Evaluation

All models were evaluated using the test set.

Model performance was assessed using standard information extraction metrics, including accuracy, precision, recall, and macro-averaged F1-score. For numerical targets, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were reported to capture prediction deviations.

Computational cost was measured as CPU runtime (seconds), CPU memory peak (in Mb), and GPU memory peak (in Mb).

This study evaluates two extraction strategies for structuring medication prescription instructions: a rule-based approach and transformer-based models. Both methods were designed to identify three target attributes dosage, duration, and medication unit from free-text prescription records.

3.2.2. Models and computational setup

All models used the variables: *prescription* (free text), *product strength* (free text) and *quantity* (numerical) as their input to extract the six target variables. While *quantity* is not free text, it is needed to arrive at the duration: duration can be calculated by dividing the prescribed quantity by the daily dosage (often present in the prescription field).

Rule-based model. The rule-based system was implemented using handcrafted patterns and regular expressions to capture structured dosage expressions. The rules were iteratively refined to accommodate common prescribing formats, including numerical ranges and abbreviated notation.

Transformer-based models. To complement the rule-based model, transformer-based language models were evaluated for their ability to generalize across linguistic variation. We employed two Dutch-language models: RobBERT, a general-domain RoBERTa-based model, and MedRoBERTa.nl, a

Variable name	Description	Type
Min quantity	Minimum quantity of the prescribed units per day	Numerical
Max quantity	Maximum quantity of the prescribed units per day	Numerical
Min duration	Minimum duration (in days) of the prescribed treatment	Numerical
Max duration	Maximum duration (in days) of the prescribed treatment	Numerical
Strict label	Whether the prescription was strict (e.g., not strict could be "as needed")	Categorical
Unit label	The unit used to indicate the strength of the medication (e.g., milligrams, milliliters)	Categorical

Table 1: Overview of the six target variables

Column	n compared	Agreement rate (%)	Cohen's Kappa κ	Mean absolute difference
Min quantity	100	87.85	0.767	0.150
Max quantity	100	86.92	0.767	0.159
Min duration	100	85.85	n/a	4.632
Max duration	100	85.85	n/a	4.934
Strict label	100	98.13	0.490	n/a
Unit label	100	87.62	0.517	n/a

Table 2: Inter-annotator agreement for 100 examples. Mean absolute differences are reported in units/day for quantity fields and in days for duration fields.

domain-specific model pre-trained on Dutch electronic health records.

Using the training set, each model was fine-tuned in both a multitask (MT) setting (sharing weights of the transformer, to jointly output all target labels at once) and a per-target (PT) setting (separate transformer weights per target label).

Models were fine-tuned using AdamW (learning rate of $1x10^-5$, and weight decay of 0.01), batch size 32, using mixed precision (PyTorch AMP), with a single Nvidia T4 GPU) for 30 epochs. Cross-entropy loss and Huber loss ($\delta = 10$) were used for categorical and numerical outcomes respectively. For categorical outcomes, to address the high class imbalance, class weights were applied during training. For the numerical targets, as the distributions were very left-skewed, a log transformation was used (inverted at evaluation).

Computational setup Experiments were conducted with the following allocated resources: AMD EPYC 7V12 with 4 vCPUs (1thread/core), Nvidia Tesla T4 with 16GB VRAM (2.7GB used during measurement), Driver 535.247.01, CUDA 12.2, a x86_64 (64-bit), Little Endian, BogoMIPS 4890, Full VM (Hyper-V).

4. Results

This section presents the performance of the rule-based and transformer-based approaches on the extraction of medication prescription target attributes. The comparison focuses on extraction performance and computational costs.

Overall, the rule-based system achieved the highest extraction accuracy on structured prescription patterns, confirming its effectiveness in scenarios where expressions closely match predefined rules. However, its performance declined when encountering unseen abbreviations or irregular dosage formats that were not covered by the handcrafted patterns. Transformer-based models demonstrated greater robustness when processing complex or composite instructions and were better able to handle off-pattern phrasing.

Both approaches showed reduced performance on specific prescription types. In particular, long treatment durations were frequently underestimated, and fine-grained unit distinctions (e.g., similar lexical forms) led to classification errors. Complex and system-generated dosage patterns were also prone to misinterpretation.

From a computational perspective, the rule-based method was substantially more efficient, requiring minimal runtime (0.02 seconds) and memory (112.7 MB) while operating without GPU resources. Transformer models, in contrast, demanded significantly greater computational capacity, particularly during training.

5. Discussion

This study provides a comparative evaluation of rule-based and transformer-based approaches for extracting structured medication information from Dutch primary care EHRs. The results demonstrate that both strategies offer distinct advantages, and their suitability largely depends on the intended deployment context.

The rule-based system delivered consistently strong performance, particularly for duration prediction and unit classification. Its deterministic design enables stable numerical extraction when prescription patterns follow recognizable structures. In addition, the minimal computational requirements make

Model	Strict label Accuracy	Precision	Recall	F1	Unit label Accuracy	Precision	Recall	F1
Rule-based	0.98	0.95	0.99	0.97	0.9	0.76	0.87	0.81
RobBERT (multitask)	0.99	0.95	0.99	0.97	0.82	0.74	0.78	0.74
MedRoBERTa (multitask)	0.99	0.97	0.99	0.99	0.79	0.76	0.73	0.71
RobBERT (per target)	0.99	0.97	0.99	0.98	0.83	0.78	0.74	0.73
MedRoBERTa (per target)	1	1	1	1	0.86	0.8	0.77	0.76

Table 3: Extraction performance on the test set for categorical targets.

Model	Min quantity RMSE	MAE	Max quantity RMSE	MAE	Min duration RMSE	MAE	Max duration RMSE	MAE
Rule-based	0.12	0.03	1.2	0.18	0.43	0.04	8.16	1.07
RobBERT (multitask)	0.29	0.36	0.38	0.33	2377.8	33.5	1386.1	25.17
MedRoBERTa (multitask)	0.2	0.28	0.29	0.27	1479.7	25	1409.4	24.23
RobBERT (per target)	0.06	0.17	0.16	0.2	4382.9	43.7	1673.4	27.3
MedRoBERTa (per target)	0.02	0.1	0.17	0.17	1692	26.3	1508.9	24.6

Table 4: Extraction performance on the test set for numerical targets.

Model	**Split**	**Runtime [s]**	**CPU peak [MB]**	**GPU peak [MB]**
Rule-based	Test	0.02	112.7	–
RobBERT (MT)	Train	312.76	1423.9	2271.2
	Test	0.10	1524.0	1134.0
MedRoBERTa (MT)	Train	176.38	1506.0	2443.4
	Test	0.11	1521.0	1203.0
RobBERT (PT)	Train	961.46	2522.9	6743.0
	Test	0.10	1894.0	1034.0
MedRoBERTa (PT)	Train	986.65	2683.8	7255.5
	Test	0.14	1521.0	1203.0

Table 5: Average runtime (CPU time in seconds) and peak memory use (MB) over 5 runs.

it especially attractive for secure research environments where external dependencies and hardware resources may be limited.

Transformer-based models, in contrast, showed greater adaptability to linguistic variability and complex prescription phrasing. However, this flexibility came at the cost of increased computational demand and reduced robustness in extracting certain numerical targets.

The observed performance differences emphasize an important trade-off between interpretability and adaptability. Rule-based systems provide transparent decision logic that facilitates error tracing and supports clinical trust, whereas transformer models operate largely as black boxes despite their predictive capabilities. In healthcare settings, where explainability is often a prerequisite for adoption, this distinction remains highly relevant.

Another key finding concerns model stability. While transformer models achieved near-perfect classification scores, they exhibited higher variance in duration predictions, particularly for extreme treatment lengths. This suggests that numerical extrapolation remains challenging for data-driven approaches trained on limited samples.

Taken together, these findings indicate that rule-based and transformer-based methods should not necessarily be viewed as competing alternatives. Instead, they may function as complementary components within hybrid extraction pipelines.

5.1. Limitations

Several limitations should be considered when interpreting these results. First, the annotated dataset is relatively small which may impact generalization, and was not split temporally, which could lead optimism in the evaluation estimates with regard to recent samples. Second, the data originate from a single regional general practitioner's network, potentially limiting portability to regions with different prescribing conventions. Finally, transformer models were evaluated without extensive data augmentation or advanced optimization strategies, leaving room for model improvements.

5.2. Implications for Clinical NLP

From a practical perspective, these results highlight the importance of aligning model selection with operational constraints. Lightweight systems may provide immediate value in resource-constrained environments, whereas transformer-based approaches become increasingly attractive as larger annotated datasets become available.

More broadly, this study contributes empirical evidence from the underrepresented domain of Dutch clinical NLP, supporting ongoing efforts to expand language resources beyond English-centric research.

6. Conclusion

This paper presented a comparative study of rule-based and transformer-based methods for extracting structured medication prescription information from Dutch primary care EHRs. The results demonstrate that rule-based extraction remains a highly reliable and computationally efficient solution for structured prescribing patterns, while transformer models provide greater flexibility in handling linguistic variation.

Rather than identifying a single superior approach, our findings highlight the complementary strengths of deterministic and data-driven methods. Selecting an appropriate strategy therefore depends not only on predictive performance but also on operational constraints such as computational resources, interpretability requirements, and deployment context.

By providing empirical evidence from the relatively underexplored domain of Dutch clinical NLP, this work contributes to the development of scalable approaches for structuring prescription data and supports the broader goal of improving data readiness for large-scale clinical research.

Future work should focus on expanding annotated datasets, improving robustness for numerical prediction tasks, and exploring hybrid architectures that combine the stability of rule-based systems with the adaptability of transformer models.

Acknowledgments

We thank Wijnand Laan for his consultations on the JGPN data processing, Nicole Boekema-Bakker for setting up and managing the digital research environment, and Sander van Doorn and Boris de Graaff for their help in the annotations.

Ethical statement

All data in this study were processed in accordance with the GDPR-compliant procedures of the JGPN, and the protocol was reviewed in consultation with a privacy officer of Utrecht University.

Data availability

Due to the privacy sensitivity of the data they could not be shared. Code used to conduct the experiments is publicly available on GitHub: https://github.com/MariaDukmak/UMC-medication-extraction-submission-/tree/main.

References

Mohamed AlShuweihi, Said A. Salloum, and Khaled Shaalan. 2021. *Biomedical Corpora and Natural Language Processing on Clinical Text in Languages Other Than English: A Systematic Review*, pages 491–509. Springer International Publishing.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2025. Hybrid natural language processing tool for semantic annotation of medical texts in spanish. *BMC bioinformatics*, 26(1):7.

Thibaut Fabacher, Erik-André Sauleau, Emmanuelle Arcay, Bineta Faye, Maxime Alter, Archia Chahard, Nathan Miraillet, Adrien Coulet, and Aurélie Névéol. 2025. Efficient extraction of medication information from clinical notes: an evaluation in 2 languages. *Journal of the American Medical Informatics Association*, 32(12):1855–1864.

E. Ford, J. Carroll, Helen E. Smith, D. Scott, and J. Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, 23:1007 – 1015.

Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2022. Information extraction from electronic medical documents: State of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516.

Hielke Muizelaar, Marcel Haas, Koert Van Dortmont, Peter Van Der Putten, and Marco Spruit. 2024. Extracting patient lifestyle characteristics from Dutch clinical text with bert models. *BMC Medical Informatics and Decision Making*, 24(1).

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923–936.

Jon Patrick and Min Li. 2009. A cascade approach to extracting medication events. *ResearchGate*.

Hugo M Smeets, Marlous F Kortekaas, Frans H Rutten, Michiel L Bots, Willem van der Kraan, Gerard Daggelders, Hanneke Smits-Pelser, Charles W Helsper, Arno W Hoes, and Niek J de Wit. 2018. Routine primary care data for scientific research, quality of care programs and educational purposes: the julius general practitioners' network (jgpn). *BMC health services research*, 18(1):735.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews Computational Statistics*, 13(6).

{Nicolas H} Thurin, Romin Pajouheshnia, Giuseppe Roberto, Caitlin Dodd, Giulia Hyeraci, Claudia Bartolini, and et al. 2022. From inception to conception: Genesis of a network to support better monitoring and communication of medication safety during pregnancy and breastfeeding. *Clinical Pharmacology and Therapeutics*, 111(1):321–331.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Stella Verkijk and Piek Vossen. 2021. MedRoBERTa.nl: A language model for Dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, 11:141–159.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Supplementary Materials

Listing 1: Label Studio template (part 1)

```
<View>
  <Header value="[Annotatie] JHN Medicatie
      Tabel (100 voorbeelden)" style="font-size:
      20px; margin: 10px 0;"/>

  <!-- Informatie uit dataset -->
  <View style="border: 2px solid #ccc; border-
      radius: 10px; padding: 15px; margin: 15px 0;
       background-color: #f9f9f9;">
    <Header value="[Informatie uit dataset]" style
      ="font-size: 16px; margin-bottom: 10px;"/>
    <Text name="id_text" value="[ID] $Id"/>
    <Text name="voorschrijfdatum_text" value="[
      Voorschrijfdatum] $Voorschrijfdatum"/>
    <Text name="einddatum_text" value="[
      Einddatum] $Einddatum"/>
    <Text name="hoeveelheid_text" value="[
      Hoeveelheid] $Hoeveelheid"/>
    <Text name="product_sterkte_text" value="[
      Product Sterkte] $Product_sterkte"/>
    <Text name="gebruik_text" value="[
      Gebruiksvoorschrift] $Gebruiksvoorschrift"/>
  </View>

  <!-- Gebruikvoorschrift Details -->
  <View style="border: 2px dashed #9ad7df;
      border-radius: 10px; padding: 16px; margin:
       10px 0 20px; background-color: #eaf7fb;">
    <Header value="[Gebruikvoorschrift Details]"
      style="font-size: 16px; margin-bottom: 8px;
      "/>
    <View style="font-size: 13px; line-height: 1.4;
       margin-bottom: 10px;">
      <Text name="instructions" value="- Min/Max
       presc per day: Beschrijf in getallen (bv. '1-2
       tablet per dag' min 1, max 2)."/>
      <Text name="instructions2" value="- Min/
      Max presc duration days: Geef cijfers op
      basis van hoeveelheid en gebruiksvoorschrift
      ."/>
      <Text name="instructions3" value="- Strict /
      Not Strict: Strict betekent dat dit nauw moet
      worden opgevolgd; anders waar 'zo nodig'
      staat of / niet strict."/>
    </View>
  </View>
</View>
```

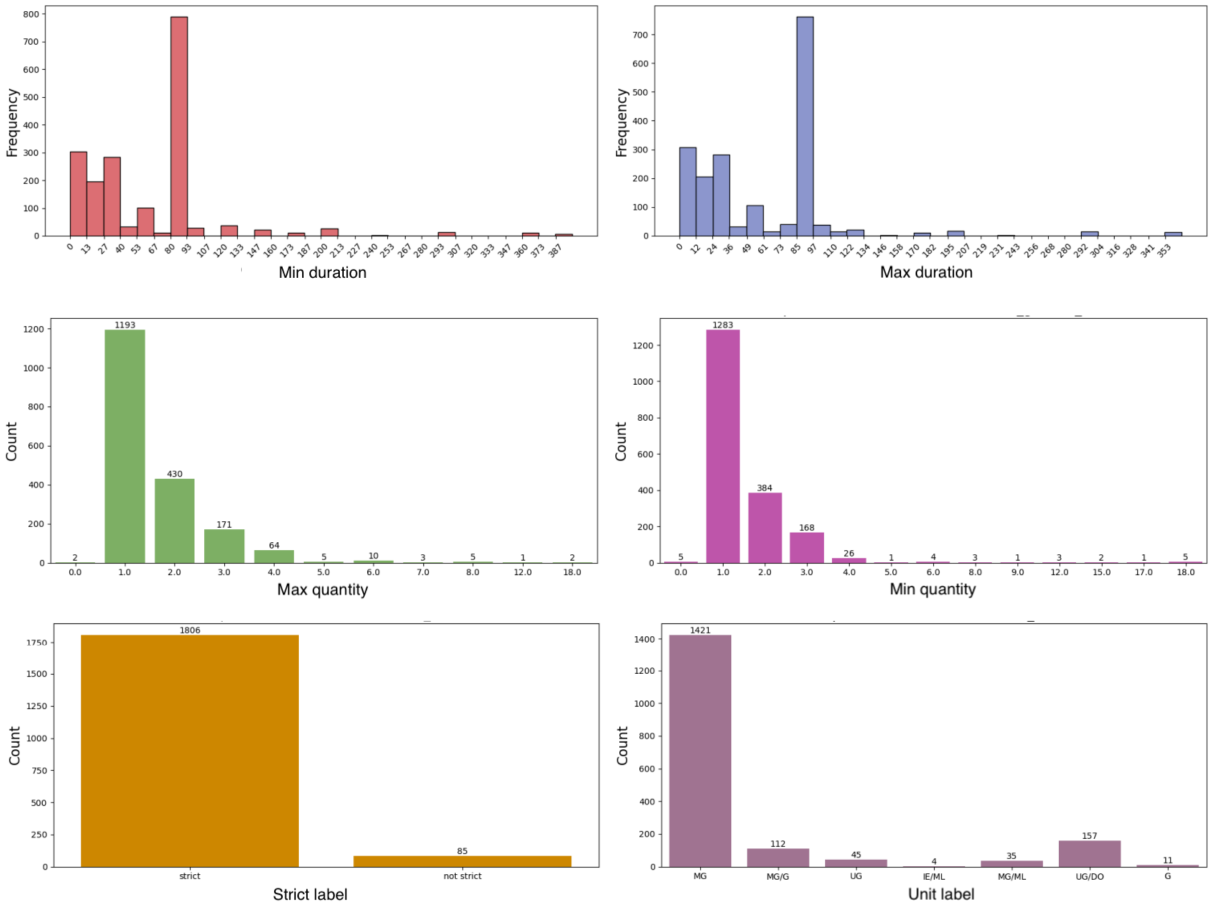

Figure 2: Visualisation of the target label distributions.

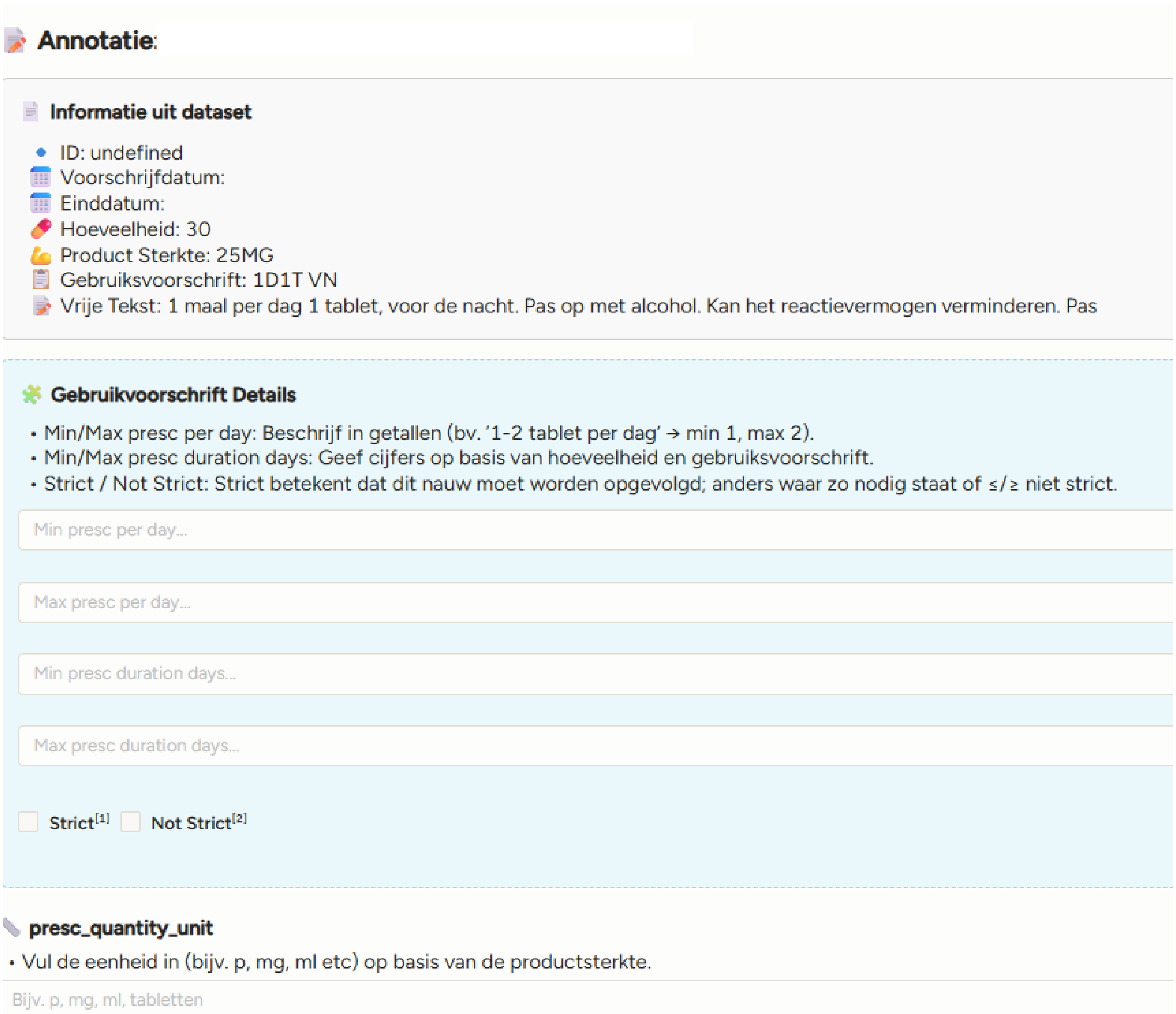

Figure 3: A screenshot of the annotation environment (for an artificial example).