# Annotating Health Records: Does Ground Truth Even Exist?

Kristof Anetta

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

**Abstract.** This paper introduces a new ground truth subset of the CSEHR dataset, a dataset of Czech health records annotated using a schema of 14 classes that is an adapted version of Apache cTAKES Core Clinical Element types. The paper details the considerations involved in (re)defining individual annotation classes in attempts to maximize utility in computational understanding of medical text.

**Keywords:** Czech, Electronic health records, EHR, annotation, named entity recognition, NER, medical concept mining.

## 1 Introduction

The development of medical entity recognition models requires an annotated health record dataset, and the annotation of a health record dataset requires a well-chosen annotation schema. This paper will present some of the considerations encountered during work on augmenting the CSEHR dataset [3], a recent annotated dataset of Czech oncology health records from the Masaryk Memorial Cancer Institute in Brno, Czech Republic.

The initial version of the CSEHR dataset involved 11 student annotators with varying degrees of precision and recall (with respect to hypothetical perfection in following the annotation manual). Given the constraints of time

Table 1: Statistics of CSEHR dataset with added ground truth.

| | Original student data | New ground truth |
|---|---|---|
| Records | 168 | 20 |
| Sentences | 3,566 | 801 |
| Words | 49,530 | 12,547 |
| Tokens | 69,699 | 19,095 |
| Percentage of tokens annotated | 40.6% | 53.5% |
| Number of tokens annotated | 28,329 | 10,218 |
| Total number of annotation tags | 31,610 | 11,556 |

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 65–74, 2024. 

and resources, the annotation quality was sufficient for the dataset to become a promising base for bootstrapping larger datasets and training larger models, and thus be treated as "truth". Nevertheless, the inherent imperfections of student annotation prevented the evaluation from being anything more than n-fold cross-validation. For a better bootstrapping protocol, we were missing a "truer truth" that would diminish the propagation of annotation faults.

Hence, our next step was to develop an addition to the CSEHR dataset, an improved section, the creation of which involved full, one-person focus on consistency and density, homogenizing annotation rule interpretations and avoiding omissions. Just like in the case of student annotation, the BRAT annotation tool [7] was used, together with automated scripts for fixing errors. The comparison of the student portion of the dataset and the newly developed ground truth can be found in Table 1.

## 2 Annotation schema

The new ground truth was annotated using the same 14-class schema as the larger student annotation section, one based on the 6 core clinical elements [1]

Table 2: 14 entity categories used for annotating the CSEHR dataset, compared with Apache cTAKES types themselves and other notable annotation schemas: Zhu et al. [10], CLEF [5], and i2b2 [8]. Dark rectangles mark the overlap of classes above and below.

| Apache cTAKES | Our annotation schema | Zhu et al. | CLEF | i2b2 |
|---|---|---|---|---|
| DiseaseDisorder | **DiseaseDisorder** | Disease | Condition | Medical Problem |
| SignSymptom | **SignSymptom** | Symptom | | |
| Medication | **Medication_name** | Drug | Drug | Treatment |
| | **Medication_strength** | | | |
| | **Medication_dosage** | | | |
| Procedure | **Procedure** | Treatment | Intervention | |
| | | | Investigation | |
| Lab | **Lab_name** | Test | | Test |
| | **Lab_value** | | Result | |
| | **Lab_unit** | | | |
| AnatomicalSite | **AnatomicalSite_name** | Body | Locus | |
| | | | Sub-location (modifier) | |
| | **AnatomicalSite_laterality** | | Laterality (modifier) | |
| | **Negation** | | Negation (modifier) | |
| | **DateTime** | | | |
| | **Abbreviation** | | | |
| | | Personal History | | |
| | | Equipment | | |
| | | Department | | |

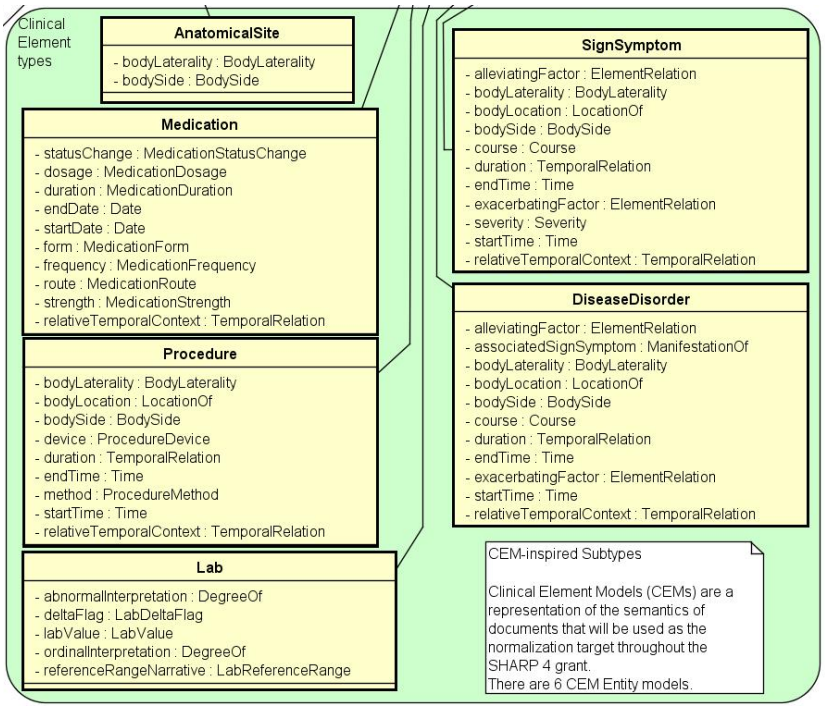



Fig. 1: Core entities in the type system of Apache cTAKES [1].

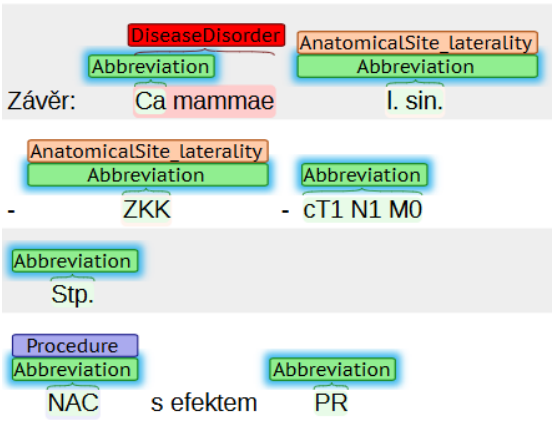
(Figure 1) in the type system of Apache cTAKES [6], an open-source NLP system for the extraction of clinical information from free text. The complete list of tags in this adapted schema can be seen in Table 2, next to comparable annotation schemas in related literature. The non-medical categories *Abbreviation*, *DateTime*, and *Negation* were added for practical reasons.

## 3 Related work in designing annotation schemas

Although the Apache cTAKES [6] type system is just one of many attempts to efficiently represent salient information in medical text, its intuitions are very similar to the efforts of other teams and organizations. Zhu et al. [10], after reviewing a number of existing medical corpora, synthesized an annotation schema of *Disease*, *Symptom*, *Test*, *Treatment*, *Drug*, and *Body*, which corresponds almost exactly to that of cTAKES, with the addition of classes for *Personal History*, *Equipment*, and *Department*.

The CLEF corpus [5] from 2007 also employed an annotation schema with significant overlaps, as shown in Table 2.

Perhaps the most well-known schema, that of the 2010 i2b2 challenge [8], has only three types (*Medical Problem*, *Treatment*, and *Test*) in its annotation guideline [2]. Although the granularity is markedly different, these three general types subsume most of the other schemas' more specific types.




```
<p>
<s>
Závěr     klgInScl závěr-n  závěr
<g/>
:         kIx,      :-x     :
Ca        kA        ca-x    ca        Carcinoma                                          B-DiseaseDisorder             B-Abbreviation
mammae    klgFnPc2  mammae-n mammae                                                     I-DiseaseDisorder
l         kA        l-x     l         vlevo                                            B-AnatomicalSite_laterality   B-Abbreviation
<g/>
.         kIx.      .-x     .         vlevo                                            I-AnatomicalSite_laterality   I-Abbreviation
sin       kA        sin-x   sin       vlevo                                            I-AnatomicalSite_laterality   I-Abbreviation
<g/>
.         kIx.      .-x     .         vlevo                                            I-AnatomicalSite_laterality   I-Abbreviation
-         kIx~      --x     -
ZKK       kA        ZKK-x   ZKK       zevní kvadranty                                  B-Abbreviation                B-AnatomicalSite_laterality
-         kIx~      --x     -
cT1       k4        cT1-m   cT1       klinická klasifikace tumor 1, nodus 1, metastáze 0   B-Abbreviation
N1        klgMnScl  N1-n     N1        klinická klasifikace tumor 1, nodus 1, metastáze 0   I-Abbreviation
M0        klgMnScl  M0-n     M0        klinická klasifikace tumor 1, nodus 1, metastáze 0   I-Abbreviation
</s>
</p>
<p>
<s>
Stp       k?        Stp-x   Stp       stav po                                          B-Abbreviation
<g/>
.         kIx.      .-x     .         stav po                                          I-Abbreviation
</s>
<s desamb="1">
NAC       kA        NAC-x   NAC       neoadjuvantní chemoterapii                       B-Abbreviation                B-Procedure
s         k7c7      s-p     s
efektem   klgInSc7  efekt-n  efekt
PR        k0        pr-i    pr        parciální remise                                 B-Abbreviation
</s>
</p>
```

Fig. 2: Example of annotation including nested annotations shown both in BRAT and in BIO format.

## 4 Considerations

### 4.1 Coverage of the schema

There are several different possible goals one can have in mind when annotating health records. They include, but are not limited to:

1. **Precise retrieval of decision-crucial classes**, such as looking for disease names.
   - Advantages: Sensitivity to a limited number of high-impact classes, easier to create training data.
   - Disadvantages: Most of the detailed information is ignored.
2. **Complete ontological knowledge of the text**, which aims to be able to categorize almost every word.
   - Advantages: Most of the text is represented on multiple levels, searchable, suitable for relation extraction.
   - Disadvantages: Very high number of classes, ambiguous classes, difficult annotation and training, lower precision.

3. **Extraction of structured information**, which puts emphasis on the discrete and continuous variables hidden in plain text.
   - Advantages: Often, structured information has well-defined orthographical forms and available values.
   - Disadvantages: There are too many possible structured information slots pertaining to a patient. Unless we know what information we are looking for, we are unlikely to get it.

Currently, the annotation schema of CSEHR cuts somewhere between 1 and 2. Its focus on 6 core clinical elements gets salient information about both the patient (*DiseaseDisorder*, *SignSymptom*, *AnatomicalSite*) and how they are treated (*Procedure*, *Medication*, *Lab*). However, the schema is missing a significant portion of health record text simply because much of the information does not fall under the 6 clinical elements. The following section details these gaps in representation.

### 4.2 Unrepresented concepts

Notably, the gaps occur in:

- Specifying adjectives.
  `jódového` (*iodine* (*adjective*))
  `vstřebatelné` (*absorbable*)
  `alveolární` (*alveolar*)
  `závažných` (*severe*)
  `objemné` (*voluminous*)
  are only a small sample of unannotated medically relevant properties of concepts.
- Nouns of medical objects and devices.
  `(jódové) zrno` ((*iodine*) *grain*)
  `stehy` (*stitches*)
  `RTG 3 GE Discovery XR 656` (*machine name*)
  all remain outside the annotation schema because, while related to a *Procedure*, they are only its means or results. In post hoc analysis, we found that an addition of a *Procedure_device* class would cover this gap and it will be considered for the next iterations of the dataset.
- Narrow domain concepts, in this case highly specialized oncological and genetic sections
  `cT2N1histolog.M0` (*cancer staging notation*)
  This spaceless blend of codes and an abbreviated word refers to clinical TNM (tumor-nodus-metastasis) staging of cancer: tumor size stage 2, nodular involvement stage 1 with histological confirmation, no metastases. The closest annotations suitable for the string might be those of the *Lab* class expanded for more complex observations, if the string were broken down into single-letter units of meaning.
  `NBN/c.657_ 661del/p.LysAsnfs*16` (*genetic mutation*)

This is a detailed specification of a genetic mutation. Its closest annotation in the schema might be *SignSymptom* as it is a description of a mutation which is an observed sign, although that does not do justice to the specific information contained therein.

### 4.3 Definitional conundra per clinical element

**AnatomicalSite** Theoretically, an anatomical site can be as big as the

`trup` (*trunk*)

and as small as a

`gen` (*gene*)

Is a

`uzlina` (*nodule*)

an *AnatomicalSite* when it can refer to any nodule in the human body? Are adjectives like

`plicní` (*pulmonary*)

sufficient references to an anatomical site to be annotated as such? In the ground truth annotation, we put the practical boundary of *AnatomicalSite* so that only body parts visible to the human eye (or visible if surgically removed) are annotated as such (hence including the nodule) and adjectives are not considered *AnatomicalSite*s.

**Lab** Interpreted strictly, the pair of *Lab_name* and *Lab_value* (*Lab_unit* optional) only refer to laboratory measurements related to the patient, such as

`P_Natrium: 143 mmol/l` (*id.*)

However, the regular orthographical structure of this class naturally offers an expansion to more general body measurements not determined in a laboratory, e.g.

`BMI: 25.28` (*id.*)
`TK/puls: 123/ 90/ 88` (*blood pressure/heart rate*)

As long as the values remain numeric, this expansion seems clearly suitable for training NER models.

The definition of *Lab_value*s as numeric might be too constraining if we want to capture a variety of observation values. The *Lab* class is especially promising for the extraction of structured information by slot filling, the *Lab_name* corresponding to a slot to be filled with *Lab_value*. Therefore, this paper will delve into a more extensive analysis.

Some laboratory results have a string for value

`HER2 negativní` (*HER2 negative*)

and, together with the non-laboratory measurements mentioned above, they open an avenue for an entire continuum of examination observations to fall into the *Lab* class. This ambiguous space stretches from fairly measurement-like entries such as

`porody: 2` (*childbirths: 2*)

to statements of more or less specified presence or absence of a condition

`thyreopatie 0` (*thyreopathy 0 = no thyreopathy*)

to discrete observations of reported behavior

`kouření: nyní` (*smoking: currently*)

At this point, if all of the above are allowed to be annotated as *Lab*, the class begins to stretch dangerously, resulting in three main issues:

1. Almost every copular clause linking a complement to a medically relevant subject becomes a candidate for the *Lab* class:
   `mamy symetrické` (*mammae symmetrical*)
   `břicho nebol.` (*abdomen non-tender*)
2. The dilemma of multiple values. Some measurements and observations yield multiple values.
   `Břicho v niveau, měkké, prohmatné, palpačně nebolestivé, bez rezistence, neperitoneální, poklep dif. bubínkový, peristaltika v normě` (*Abdomen at level, soft, palpable, non-tender on palpation, without resistance, non-peritoneal, percussion diffusely tympanic, peristalsis normal*)
   When should such a structure stop being treated as a name-value(s) pair? In our annotation schema, as long as the list of observations after a *Lab_name* is contiguous, they may all be annotated as *Lab_value*s.
3. In instances of hierarchically structured observations such as
   `Hlava : poklepově nebolestivá, výstupy trigeminu nebolest., zornice izokorické` (*Head: non-tender to percussion, trigeminal nerve exits non-tender, pupils isocoric*)
   in one perspective,
   `Hlava` (*Head*)
   is the *Lab_name* meaning "examination of the head" and all items that follow are *Lab_value*s, as they are all observations related to the head. At the same time,
   `zornice` (*pupils*)
   is an even more precise *Lab_name*, meaning "status of pupils during head examination"
   `izokorické` (*isocoric*)
   being its *Lab_value*. Hence,
   `zornice` (*pupils*)
   would be both a *Lab_name* and a part of a greater *Lab_value*, and
   `izokorické` (*isocoric*)
   would be *Lab_value* twice, once completely and once partially.

***DiseaseDisorder*** In cases like

> `středně diferencovaný duktalni invazivni karcinom prsni zlazy s metastatickym postižením axillarni lymfaticke uzliny` (*moderately differentiated ductal invasive carcinoma of the breast with metastatic involvement of the axillary lymph node*)

it is difficult to determine the span of the *DiseaseDisorder* annotation. In both directions starting from

> `karcinom` (*carcinoma*)

the words are adding specification, but the more specification is added, the harder it becomes to use the whole cluster for NER model training.

For the needs of ontological representation, a reasonable solution is to find the most specific corresponding entry in the ICD [9]. For language model training, the preferred way would be that of splitting up the concept into core disease identifier and attached properties.

***Medication*** While searching for literal occurrences of medication names, the *Medication_name* class is one of the easiest to annotate - reliable databases of medications exist even in countries without other medical vocabularies.

However, the question is if the class should include concepts

- broader than a specific medicine: names of medication classes such as
  `diuretika a Ca blokátoru` (*diuretic and calcium channel blocker*)
  or even
  `medikace` (*medication*)
  itself,
- or narrower: names of active substances or chemical constituents in the medication.

## 5 Discussion

Annotated health records are a valuable resource in healthcare computing, but they are most valuable when created with a clear purpose reflected in the annotation schema. To answer the question from the title: there is never a single Medical Annotation Truth, but rather many local, purpose-bound utilitarian truths.

Though a good benchmark, the Apache cTAKES Core Clinical Element schema is designed with a realistic industry application in mind: to cherry-pick prominent occurrences of unambiguous concepts with high accuracy and link them directly to a UMLS (Unified Medical Language System) [4] dictionary entry. In the case of the *Lab* class, to extract measurement values in sufficiently regular form.

Researchers interested in finding an annotation schema to make large datasets searchable and filterable by the occurrence of key medical concepts

need look no further - the *DiseaseDisorder*, *SignSymptom*, *Procedure* and *Medication* alone can be expected to satisfy an overwhelming majority of search queries.

For researchers interested in a complete medical-ontological representation of texts and at the same time a class count reasonably low for model training, a new annotation schema should be developed, aiming specifically at high coverage and low class count. If the cTAKES types were to be adapted for this purpose, they would require some broadening (e.g. *Lab* to *Observation*) and a small number of new classes that cover medical contexts and body properties that remained outside the schema in this paper, perhaps even a class to mark narrow-domain information.

## 6 Conclusion

In this paper, we introduced the new ground truth subset of CSEHR and presented examples of decision making during annotation, especially considerations related to class definitions, class boundaries, and gaps in representation. The unexpected frequency of encountered ambiguities and gaps is a strong reminder of the need of careful planning and utilitarian class design in medical text annotation for language model training.

**Acknowledgements.** The analyzed Czech data was kindly provided by the Masaryk Memorial Cancer Institute in Brno, Czech Republic.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

## References

1. Apache cTAKES - User FAQs — svn.apache.org. https://svn.apache.org/repos/infra/websites/production/ctakes/content/user-faqs.html, [Accessed 09-11-2023]
2. Concept annotation guidelines, https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf
3. Anetta, K., Horák, A.: New human-annotated dataset of czech health records for training medical concept recognition models. In: International Conference on Text, Speech, and Dialogue. pp. 110–120. Springer (2024)
4. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research **32**(suppl_1), D267–D270 (01 2004). https://doi.org/10.1093/nar/gkh061, https://doi.org/10.1093/nar/gkh061
5. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J.S., Roberts, I., Setzer, A., Tapuria, A., et al.: The clef corpus: semantic annotation of clinical text. In: AMIA Annual Symposium Proceedings. vol. 2007, p. 625. American Medical Informatics Association (2007)
6. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)

7. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107 (2012)
8. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (2011)
9. World Health Organization: ICD-10 : International statistical classification of diseases and related health problems : Tenth revision (2004)
10. Zhu, E., Sheng, Q., Yang, H., Liu, Y., Cai, T., Li, J.: A unified framework of medical information annotation and extraction for chinese clinical text. Artificial Intelligence in Medicine **142**, 102573 (2023)