A Meta-dataset of German Medical Corpora: Harmonization of Annotations and Cross-corpus NER Evaluation

Ignacio Llorca[1], **Florian Borchert**[2], **Matthieu-P. Schapranow**[2]
Hasso Plattner Institute, University of Potsdam, Germany
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam
[1] llorcarodriguez@uni-potsdam.de
[2] {firstname.lastname}@hpi.de

Abstract

Over the last years, an increasing number of publicly available, semantically annotated medical corpora have been released for the German language. While their annotations cover comparable semantic classes, the synergies of such efforts have not been explored, yet. This is due to substantial differences in the data schemas (syntax) and annotated entities (semantics), which hinder the creation of common meta-datasets. For instance, it is unclear whether named entity recognition (NER) taggers trained on one or more of such datasets are useful to detect entities in any of the other datasets. In this work, we create harmonized versions of German medical corpora using the BIGBIO framework, and make them available to the community. Using these as a meta-dataset, we perform a series of cross-corpus evaluation experiments on two settings of aligned labels. These consist in fine-tuning various pre-trained Transformers on different combinations of training sets, and testing them against each dataset separately. We find that a) trained NER models generalize poorly, with F_1 scores dropping approx. 20 pp. on unseen test data, and b) current pre-trained Transformer models for the German language do not systematically alleviate this issue. However, our results suggest that models benefit from additional training corpora in most cases, even if these belong to different medical fields or text genres.

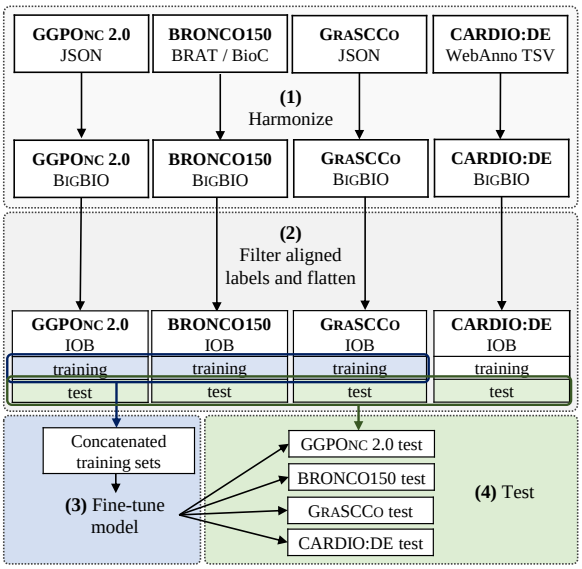


Figure 1: Overview of our experimental design. (1) Each corpus is harmonized from its source format to BIGBIO via custom schema parsers. (2) Equivalent entity classes and spans are aligned, the text is tokenized and transformed into IOB format. (3) Training splits of different corpora are concatenated (in the example GGPONC 2.0, BRONCO150 and GRASCCO) and used to train a Transformer-based NER model. (4) The model is evaluated on the test splits of all individual datasets.

1 Introduction

Recently, an increasing amount of medical text datasets for the German language with semantic annotations has been released to the public (Zesch and Bewersdorff, 2022). These corpora come in unequal data formats and with widely varying definitions of annotated entities, e.g., based on ontologies like the UMLS (Bodenreider, 2004), top level hierarchies in SNOMED CT (Donnelly, 2006), or other medical terminologies such as ICD-10. The employed annotation guidelines have usually been created ad-hoc and are hardly reused across annotation projects. Corpora are difficult to compare due to these semantic and syntactic differences. Although various NER models have been trained and tested on individual datasets, their performance across medical fields and text genres has not been investigated. Our work integrates the following distributable, annotated German corpora: BRONCO150 (Kittner et al., 2021), GGPONC 2.0 (Borchert et al., 2022), CARDIO:DE (Richter-Pechanski et al., 2023), and GRASCCO (Modersohn et al., 2022). While the latter did not contain human annotations upon release, it was recently annotated according to the GGPONC 2.0 guidelines (Bressem et al., 2023).

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 171–181

At the same time, adapted German versions of widely used Transformer models have become publicly available (Chan et al., 2020; Scheible et al., 2020), more recently also specific to biomedical texts (Lentzen et al., 2022; Bressem et al., 2023). While these models have been evaluated on many individual datasets, their performance on truly unseen data remains unknown.

To enable cross-corpus evaluations, we create a *meta-dataset* of German medical corpora by harmonizing them under the schema proposed in the BIGBIO framework (Fries et al., 2022), as illustrated in Fig. 1. This way, we also make the datasets easily available to the community as data loaders in the Hugging Face library (Llorca, 2023). An example for this schema is depicted in Fig. 2. While schema harmonization addresses the issue of syntactic interoperability, the semantics of annotated entities may still differ, as definitions of entity classes have been derived from different medical ontologies. Therefore, we propose two possible alignments of labels across the four German medical corpora and conduct a series of experiments, evaluating different combinations of training corpora and pre-trained Transformers. In this work, we focus on the entity annotations related to *medications*, as these are the only ones that can be aligned consistently across all corpora. Their definitions are still partly extracted from different medical ontologies.

Our goal is to determine whether multiple datasets with similar annotated entities contribute to creating models that can be used outside the domain they were trained in. Here we refer to the domain of a corpus broadly as the set of characteristics conforming it, mainly text genre, medical field, and annotation policy. Such experiments have been successfully conducted for the English language and resulted in robust NER taggers, albeit for entity classes different from the more clinically motivated German-language corpora (Weber et al., 2021). To the best of our knowledge, no such harmonization has been conducted for the German language and clinical entity classes.

The remainder of this work is organized as follows: in Section 2, we review literature on German clinical corpora and biomedical data harmonization. In Section 3, we describe our methods, i.e. the corpora used, data harmonization steps, and performed cross-corpus evaluation. We present our results in Section 4 and discuss them in Section 5. Our work concludes with an outlook in Section 6.

```
{'id': '0',
 'document_id': '00_mundhoehlenkarzinom_0000',
 'passages': [
   {'id': '0-0',
    'type': 'sentence',
    'text': ['Tabakkonsum ist ein wesentlicher...'],
    'offsets': [[0, 90]]}
   ],
 'entities': [
   {'id': '0-0',
    'type': 'Other_Finding',
    'text': ['Tabakkonsum'],
    'offsets': [[0, 11]],
    'normalized': []},
   {'id': '0-1',
    'type': 'Other_Finding',
    'text': ['Risikofaktor für die Entwicklung...'],
    'offsets': [[33, 89]],
    'normalized': []}
   ],
 'events': [],
 'coreferences': [],
 'relations': []}
```

Figure 2: Sample of the target schema for knowledge base construction tasks like NER from BIGBIO.

2 Related Work

In the following, we set our contribution in the context of related work.

2.1 German-language Medical Corpora

In the past, German medical text datasets have been created in closed research environments without the chance of being shared with other researchers. Notable examples include the work of Roller et al. (2016) using clinical notes from nephrology, Hahn et al. (2018) using discharge summaries of internistic or ICU units stays, and König et al. (2019) using discharge letters from the Berlin Aging Study II. Distributable corpora became available just recently, the JSYNCC corpus (Lohr et al., 2018) being a first successful example, although without semantic annotations. The BRONCO150 (Kittner et al., 2021) and CARDIO:DE (Richter-Pechanski et al., 2023) corpora are currently the only instances of annotated, distributable corpora of anonymized patient-level clinical texts. Other open corpora are based on information unrelated to individual patients, e.g. clinical guidelines in Borchert et al. (2020), or are translated versions of a public English dataset, e.g. Frei and Kramer (2022).

To the best of our knowledge, no major effort has been made in cross-corpus evaluation to assess the robustness of German biomedical NER taggers. The baseline models presented by Borchert et al. (2022) or Kittner et al. (2021) are constrained to in-domain, i.e., internal validation. Roller et al. (2022)

conducted external validation of their model on a small subset from GGPONC, but using their original annotation policy. Only Frei and Kramer (2023) and Richter-Pechanski et al. (2023) briefly report on evaluating the baseline NER model from GGPONC 2.0 on aligned medication classes from their respective datasets, with mixed results. Extending this line of research, we consider multiple possible label alignments, analyze span-wise metrics, and explore several combinations of training corpora and pre-trained Transformer models.

2.2 Data Harmonization in Clinical NLP

Several prior works have considered cross-corpus evaluations through the alignment of semantic classes across datasets for different machine learning tasks. For instance, some papers have been released on acoustic emotion recognition (Schuller et al., 2010; Zhang et al., 2011) or general NER (Nothman et al., 2009). In clinical NLP, curated dataset collections with common schemas are not frequent. Efforts like HunFlair include 31 corpora, but limited to the English language (Weber et al., 2021). There are extensive cross-corpus studies of biomedical NER models with similar entity classes and corpora, but also only for the English language (Kaewphan et al., 2016; Giorgi and Bader, 2019; Galea et al., 2018).

In relation to data standards and schemas, many annotated corpora are simply distributed in the raw format of the respective annotation tool, e.g., BRAT (Stenetorp et al., 2012) or WebAnno (Yimam et al., 2013). While formats like BioC (Comeau et al., 2013) present an attempt to standardize annotations and other metadata for biomedical text datasets, the semantics of entity annotations are not fully defined inside the standard. This is counterproductive for cross-corpus integration, as pre-processing efforts are still needed to homogenize the data.

To alleviate these problems, Fries et al. (2022) propose the BIGBIO framework, introducing fixed data schemas for different NLP tasks. BIGBIO makes minimal assumptions on pre-processing decisions to suit different sorts of datasets. In addition, it provides parsers to harmonize more than 126 corpora within this schema and allows easy access to them through the widely used Hugging Face datasets library. However, parsers for the German corpora used in this work were previously not available. Therefore, we have contributed such implementations as part of this work (Llorca, 2023).

3 Materials and Methods

In the following, we present the characteristics of each corpus and an overview of the harmonization and annotation alignment processes. We provide a description of the experimental setup and the evaluation methods used to analyze the results.

3.1 Datasets

An overview of the key details of the corpora used in our cross-corpus experiments is given in Table 1. All considered corpora have been manually annotated by medically trained personnel. Further insights on annotation policies and Inter Annotator Agreement (IAA) are given below:

- **BRONCO150**: De-identified discharge summaries annotated in two groups (A and B) of medical experts and students. IAA as micro-averaged phrase-level F_1 score ranges across entities from 0.81 to 0.94 for group A and from 0.66 to 0.87 for group B. Each semantic class is based on a different medical terminology, which are also used for grounding.
- **GGPONC 2.0**: Clinical guidelines annotated by seven medical students and curated by a medical doctor. Mean IAA, measured through the γ-method (Mathet et al., 2015), is 0.94 across all entity classes on a set of seed documents after iterative annotation guide refinement. Semantic classes are based on SNOMED CT top-level hierarchies.
- **GRASCCO**: Synthetic case reports, originally without annotations. For the benchmarks introduced by Bressem et al. (2023), it was annotated by a single medical student from the GGPONC 2.0 annotation team, following the same guidelines. Thus, the labeled entities and annotation policy are the same for both corpora. However, there is no data on annotation quality and IAA.
- **CARDIO:DE**: De-identified discharge summaries annotated by four medical informatics and two advanced medical students. Fine-grained medication information are annotated following the policy proposed by Uzuner et al. (2010). IAA is reported using token-level median F_1 scores, ranging from 0.33 to 0.98 across classes on seed documents after iterative annotation guide refinement. The lowest IAA for entity classes that we use in this work is 0.76 (*active ingredient*).

Corpus	Med. Field	Text Genre	Tokens	Format	Entities
BRONCO150 Kittner et al. (2021)	Oncology	Discharge Summaries	71K	BioC / BRAT	Diagnosis (ICD-10) Treatment (OPS) Medication (ATC)
GGPONC 2.0 Borchert et al. (2022)	Oncology	Clinical Guidelines	1,877K	JSON	Finding Substance Procedure
GRASCCO Modersohn et al. (2022)	Various	Synthetic Case Reports	43K	JSON	see GGPONC 2.0
CARDIO:DE Richter-Pechanski et al. (2023)	Cardiology	Discharge Summaries	800K	WebAnno TSV	Medications (as in Uzuner et al. (2010))

Table 1: Overview of the used corpora with their annotated entities, medical fields, text genres, size and data formats. A full list of fine-grained entity classes for each corpus can be found in Table 2.

3.2 Harmonization and Label Alignment

For each corpus, we implement a parser within the BIGBIO framework to derive a common notion of documents, passages and entity spans as outlined in Fig. 2. In order to preserve the source integrity, we consider individual sentences as the main units for our experiments, since the definitions of documents and passages differ across corpora.

To obtain semantically equivalent entity classes, we also need to align entity definitions inspired by different medical ontologies across corpora. Our attempt to do this is shown in Table 2. For GGPONC 2.0 and GRASCCO, we consider their fine-grained configuration of entity classes. In some cases, there is no exact equivalence, e.g. it is not immediately clear if *Diagnostic Procedure* in GGPONC 2.0 corresponds to *Treatment* in BRONCO150. Inspection of the annotations shows that these two do not overlap fully, unlike *Therapeutic Procedure* and *Treatment*. Therefore, *Diagnostic Procedure* is left unmapped.

Medications are the only entity class that can be consistently found across all corpora, although its definition is not identical. In fact, CARDIO:DE contains only medication annotations, but much more fine-grained than in the other corpora. BRONCO150 annotations leave out the dosage information of a medication, while CARDIO:DE annotations consider it with dedicated labels. GGPONC 2.0 (and GRASSCO) offer two span length configurations: the short configuration matches the BRONCO150 definition, while the long one covers the *Strength* and *Frequency* annotations from CARDIO:DE as well. Therefore, we can align annotated spans across all corpora as shown in Table 3.

Cases where several medication annotations are either nested or overlap are not possible in some corpora and very seldom in others. Thus, the loss of information when flattening the datasets into IOB format is minimal. Non-contiguous annotations are treated as separate entities, following the same principle used for the NER models in the papers from BRONCO150 and GGPONC 2.0.

3.3 Cross-Corpus Evaluation Experiments

As a result of the above assumptions, we only consider annotations of medication entities for our cross-corpus NER evaluation. We use the following configurations of label alignments:

- **Short-span**: Short-span version of *Clinical Drug* entities from GGPONC 2.0 and GRASCCO, the *Medication* annotations from BRONCO150 and *Drug / Active Ingredients* from CARDIO:DE (discarding linked *Strength* and *Frequency* annotations), resulting in 15 combinations of training corpora.
- **Long-span**: Long-span version of GGPONC 2.0 and GRASCCO, discarding BRONCO150, and merging *Drugs*, *Strength*, and *Frequency* annotations from CARDIO:DE that are linked to each other, as in Richter-Pechanski et al. (2023), resulting in seven combinations.

Afterward, we perform two sets of experiments:

(i) In a larger set of experiments, we fine-tune a Transformer model with a token classification head on all combinations of training data, and evaluate it separately against the test split of each corpus. For these experiments, we use

GGPONC 2.0, GRASCCO	BRONCO150	CARDIO:DE
Diagnosis / Pathology	Diagnosis	–
Clinical Drug	Medication	Active Ingredient, Drug
Therapeutic Procedure	Treatment	–
Other Finding, Diagnostic Procedure, Nutrient / Body Subst., External Subst.	–	–
–	–	Dosage, Route, Form, Reason, Duration, Strength, Frequency

Table 2: Mapping of annotated semantic classes for named entities across datasets. (–) indicates that there are no entities in a certain corpus equivalent to the entity of other dataset. Only the semantic classes for medications (Clinical Drug, Active Ingredient, Medication, Drug) can be mapped across all four corpora.

Example	`Metroprolol`	`95 mg`	`1-0-1`
GGPONC (L)	Clinical Drug		
GGPONC (S)	Clinical Drug	O	
BRONCO150	Medication	O	
CARDIO:DE	Active Ing.	Strength	Freq.

Table 3: Example of how the annotation policies for medications vary in each corpus and how they can be aligned. For GGPONC, L and S refer to the long and short configurations. These apply equivalently for GRASCCO.

the recent BioGottBERT (Lentzen et al., 2022) as the pre-trained Transformer.

(ii) In a second set of experiments, we compare the impact of different Transformer checkpoints on the out-of-domain robustness of trained NER models. For this purpose, we consider only the long-span combinations with two training datasets and an unseen test dataset. The models we compare are GBERT and GELECTRA (Chan et al., 2020), BioGottBERT (Lentzen et al., 2022), and medBERT.de (Bressem et al., 2023).

Despite BRONCO150 having five splits for cross-validation, incorporating this would greatly increase the complexity and number of experiments. Instead, we separate one random split for testing. Similarly, CARDIO:DE does not have pre-defined splits. Thus, we randomly sample a validation and test set containing 12.5 % of all documents, fixed for all experiments.

As hyperparameters, we use a learning rate of 5×10^{-5}, with linear decrease and no weight decay, warmup or label smoothing. All models are trained for 50 epochs on a single NVIDIA A40 GPU with a batch size of 32.

3.4 Evaluation Metrics

We make use of two evaluation methods: seqeval and FairEval. Seqeval is widely used in the field for sequence labeling evaluation and provides a traditional F_1 score implementation (Nakayama, 2018). FairEval is a novel approach to subdue the double-penalties that occur in traditional evaluation when a prediction misses the boundaries of an annotation (Ortmann, 2022). It also provides more fine-grained metrics for error analysis, as it outputs *true positives* (TP) and separates *boundary errors* (BE) from *false positives* (FP) and *false negatives* (FN). In order to ease its usability for the community, we implemented FairEval as a publicly available Hugging Face evaluation module (Llorca, 2022). For the aggregation of scores across test sets, we follow the conclusions of Forman and Scholz (2010) and give greater importance to the micro-averaged results. Macro scores are still reported, accounting for the large size imbalance among the datasets.

4 Results

The seqeval (traditional) F_1 scores of the first set of experiments are shown in Table 4 and 5 for the short and long-span setting, respectively. We omit FairEval scores for this set of experiments for brevity, as the directionality of results is the same.

We use abbreviations with the first three letters to refer to the datasets, i.e. GRASCCO is GRA. We recall the experiments by the row number in Table 4 and 5 or by the following notation: BRO+GGP→BRO corresponds to the model trained on BRONCO150 and GGPONC 2.0, and tested on BRONCO150, i.e. the first cell in Table 4, row 7 (0.925).

4.1 Out-of-domain Generalization of Clinical NER Models

In general, models perform considerably worse when evaluated outside their training domain, i.e., when the training split from the target corpus is not included in the joint training set. On average, F_1 scores are approx. 20 pp. points lower on unseen target corpora for both long-span and short-span experiments. The differences are especially large for the target corpus GGPONC 2.0 test set, having reductions in F_1 score of around 30 pp.

Differences are smaller when the target corpus is GRASCCO, having a decrease of around 10 pp. Notably, the short-span model trained only on GRASCCO obtains an F_1 score of 0.821 on its own test split (Table 4, row 15), while the model trained on GGPONC 2.0 alone performs just 3 pp. worse, achieving an F_1 score of 0.788 (row 14).

4.2 Effect of Adding More Training Data

In general, models are not adversely affected or mislead by adding datasets from different medical fields or text genres other than the training corpus. There are many cases where adding data from a different domain slightly improves the performance. For instance, CAR $\rightarrow$ CAR achieves an F_1 score of 0.876 (Table 4, row 13), while CAR+GGP $\rightarrow$ CAR (row 9) scores slightly higher with 0.880. The same holds true for the long-span setting: CAR+GGP $\rightarrow$ CAR outperforms CAR $\rightarrow$ CAR (Table 5, rows 17/20) by a small margin.

Cases where adding more data is only slightly detrimental are consistent across all experiments. Considering the short-span experiments with GGPONC 2.0 as the target corpus, we see how training just on itself achieves 0.910 F_1 score (Table 4, row 14) and adding more corpora decreases performance slightly up to 0.905 (for all four datasets, row 1). This finding can be observed across all experimental settings.

Such marginal loss of performance trades off positively with the robustness of models across multiple corpora. The results of the model trained on all corpora (BRO+CAR+GGP+GRA in Table 4, row 1) are slightly below those obtained by models trained on each corpus separately (shaded diagonal in Table 4, rows 12-15), while increasing the micro F_1 by a wide margin of 19 pp. on average. The same holds true for the long-span setting, with an average increase in micro F_1 of 18 pp.

4.3 Performance of Different Transformer Checkpoints

Results from the second set of experiments to investigate the impact of different pre-trained Transformer checkpoints on the out-of-domain robustness of NER taggers are presented in Table 6. This time, FairEval F_1 scores are shown together with the seqeval (traditional) scores, to gain more insights into the actual magnitude of the performance drop compared to the in-domain baseline.

For the setting tested on GGPONC 2.0, the best Transformer checkpoint varies when boundary errors are counted once instead of twice: medBERT.de obtains a higher FairEval score than BioGottBERT, whilst achieving a lower seqeval score.

There is no clear pattern with regard to the generalization capabilities of different pre-trained Transformers. GELECTRA performs best in two out of three scenarios, but falls in third place for the remaining case, where GGPONC 2.0 is the unseen target. Additionally, BioGottBERT is always the second-best checkpoint whenever GELECTRA gets the first place. The best performing Transformer for the settings tested on GGPONC 2.0 and CARDIO:DE are still far from a baseline where the model has seen the training split of the target corpus in training. In contrast, for the setting tested on GRASCCO, GELECTRA obtains a traditional F_1 score just 1 pp. below the baseline result from BioGottBERT on GRA $\rightarrow$ GRA.

5 Discussion

In this section, we discuss our findings and perform a fine-grained error analysis.

5.1 Cross-Corpus Evaluation

In general, all models perform poorly on truly unseen data, no matter if the datasets belong to the same medical field (BRONCO150 and GGPONC 2.0 concern oncology), if the annotation procedure and source format are the same (for GGPONC 2.0 and GRASCCO) or if the text genre is similar (BRONCO150 and CARDIO:DE contain discharge summaries).

When the model has not seen the target corpus during training, it performs significantly below par, which we attribute to the widely different entity definitions and annotation policies. This is the case even for a seemingly well-defined semantic class like medications. Although the pattern is less evident for the short-span configuration, this is likely

		Test set				F_1 (Average)	
		BRO	**CAR**	**GGP**	**GRA**	**Micro**	**Macro**
1	BRO CAR GGP GRA	0.928	0.876	0.905	0.783	**0.898**	**0.874**
2	BRO CAR GGP	0.930	0.873	0.906	0.694	0.897	0.852
3	BRO CAR GRA	0.916	0.876	0.593	0.776	0.694	0.792
4	BRO GGP GRA	0.937	0.708	0.906	**0.854**	0.850	0.855
5	CAR GGP GRA	0.820	0.879	0.906	0.717	0.890	0.832
6	BRO CAR	0.932	**0.883**	0.549	0.719	0.672	0.774
7	BRO GGP	0.925	0.728	0.907	0.745	0.855	0.829
8	BRO GRA	0.946	0.713	0.631	0.796	0.680	0.781
9	CAR GGP	0.812	0.880	0.901	0.694	0.885	0.823
10	CAR GRA	0.779	0.879	0.588	0.778	0.696	0.767
11	GGP GRA	0.798	0.724	0.907	0.846	0.846	0.823
12	BRO	**0.956**	0.740	0.562	0.681	0.647	0.745
13	CAR	0.754	0.876	0.489	0.687	0.628	0.708
14	GGP	0.758	0.684	**0.910**	0.788	0.834	0.786
15	GRA	0.774	0.812	0.669	0.821	0.718	0.773
	Mean on seen data	0.934	0.878	0.906	0.796		
	Mean on unseen data	0.785	0.730	0.583	0.715		

Table 4: F_1 scores (short-span setting) resulting from tuning BioGottBERT on each combination of training sets against each separate target corpus and their micro and macro aggregation. The example from Fig. 1 would correspond with row number 4. We highlight in **bold** and underlined the highest and second-highest scores for each test set. The shaded cells denote experiments where the training portion of the test corpus is seen at training. We see that (1) models generalize poorly to other domains (unshaded cells are consistently lower scores than shaded ones) and (2) models generally benefit from adding more corpora at training to the target corpus.

		Test set			F_1 (Average)	
		CAR	**GGP**	**GRA**	**Micro**	**Macro**
16	CAR GGP GRA	0.796	0.788	0.549	0.769	**0.716**
17	CAR GGP	**0.807**	0.788	0.485	**0.774**	0.698
18	CAR GRA	0.801	0.480	0.409	0.577	0.569
19	GGP GRA	0.579	**0.794**	**0.625**	0.710	0.676
20	CAR	0.804	0.424	0.258	0.543	0.504
21	GGP	0.560	0.793	0.547	0.703	0.639
22	GRA	0.593	0.496	0.606	0.532	0.599
	Mean on seen data	0.802	0.791	0.547		
	Mean on unseen data	0.577	0.467	0.430		

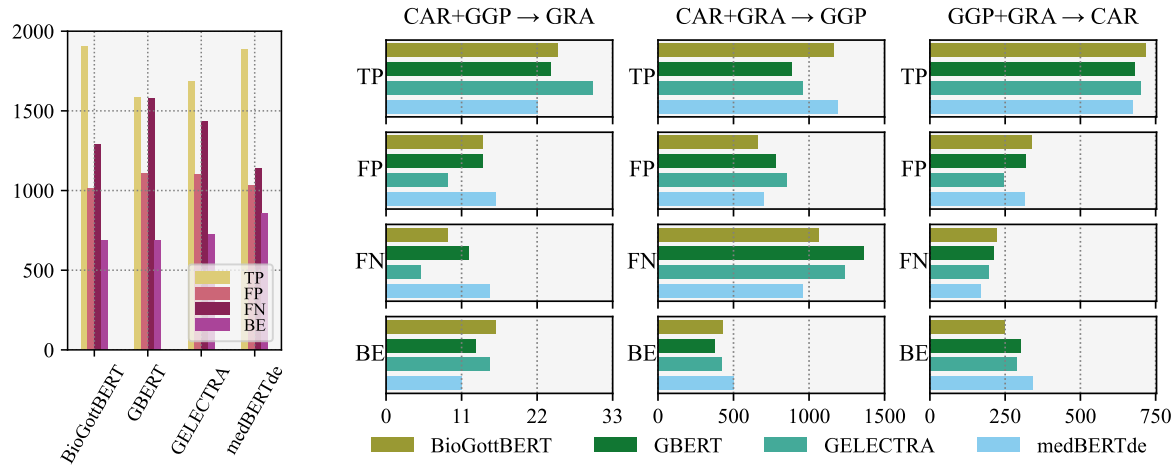
Table 5: F_1 scores (long-span setting) resulting from tuning BioGottBERT on each combination of training sets against each separate target corpus and their micro and macro aggregation. Values highlighted as in Table 4. The findings drawn in Table 4 are even more notable in this setting.

just because the task at hand is easier (i.e. shorter spans are easier to identify) – without seeing the test corpus at training, the scores remain on levels that may be deemed acceptable, but are still considerably worse.

Having models achieve higher micro and macro F_1 scores across all target sets when they have seen more corpora at training is consistent with our assumptions. The fact that adding corpora to the training split of the target corpus does not significantly reduce performance has promising implications: if there were enough open datasets, current day neural network architectures are indeed enough to obtain robust NER taggers through their combination. Concerning the cases where adding more corpora to the target corpus at training time increases model performance on it, GGPONC 2.0 seems to be the most contributing dataset: the results of CAR →

	CAR+GGP→GRA		CAR+GRA→GGP		GGP+GRA→CAR	
	seqeval	FairEval	seqeval	FairEval	seqeval	FairEval
BioGottBERT	0.485	0.562	**0.480**	0.520	0.579	0.640
GBERT	0.475	0.552	0.384	0.413	0.554	0.621
GELECTRA	**0.594**	**0.674**	0.398	0.434	**0.581**	**0.658**
medBERT.de	0.458	0.512	0.456	**0.524**	0.550	0.620
Baseline	0.606	0.723	0.793	0.839	0.807	0.846

Table 6: Out-of-domain evaluation of different Transformer checkpoints. We consider the experiments of the long-span configuration that included two corpora for training and were tested on the unseen dataset. We report traditional (seqeval) and FairEval F_1 scores to account for the effect that double penalties on close-to-target predictions have in model selection. For reference, we include the single-corpus, in-domain results achieved by BioGottBERT as a baseline (GRA → GRA, GGP → GGP, CAR → CAR).



(a) Aggregated error counts

(b) Error counts grouped by experiment and error type

Figure 3: Error counts (**T**rue **P**ositives, **F**alse **P**ositives, **F**alse **N**egatives and **B**oundary **E**rrors) per Transformer checkpoint for long-span experiments using two corpora for training and evaluated on the remaining, unseen corpus.

CAR and GRA → GRA improve when adding GGP to the training sets in both span-length configurations. Suspected reasons could be its large size, thematic diversity, or relatively high IAA. It also suggests that non-patient-related data (like clinical guidelines) can be useful to robust models when evaluated on patient-related data such as discharge summaries.

5.2 Error Analysis

The comparison of different checkpoints is initially favorable to GELECTRA, performing best in two out of three settings. It should also be noted that the best models in the last case (BioGottBERT and medBERT.de) included unlabelled texts from GGPONC in their pre-training phase

A more detailed error analysis shows that BioGottBERT and medBERT.de obtain more TPs, while producing fewer FNs and FPs aggregated through all three experiments than GELECTRA (see Fig. 3a). Furthermore, BioGottBERT also produces less boundary errors, making a case for the current most robust model on unseen data. However, the averaged trend is not consistently reflected across individual experiments (see Fig. 3b).

It is also noteworthy that general-domain models are more prone to FNs, i.e., completely missing some entities. We suppose that the reason for this is that biomedical-tuned models are more familiar with the medical terminology in the datasets. In contrast, the number of FPs is closer for all Transformer models.

6 Conclusion and Outlook

Reaching agreements and establishing standards for clinical entity annotation is vital to facilitate inter-corpus operability. So it is adhering to similar formats and schemas to structure the informa-

tion. This can help to combine the sporadically released open medical text datasets for non-English languages to build more robust models. Our work aimed to support this goal by harmonizing all currently available, semantically annotated German medical corpora within the BIGBIO framework and make the implemented data loaders available to the community.

Our experiments show that the currently available corpora in German serve for poorly generalizable models. Our results also suggest that mixing multiple corpora in training is beneficial on single test splits and widely improves robustness, thus highlighting the importance of easing cross-corpus integration. The comparison of different pre-trained Transformers does not shed conclusive results. Both general-purpose and biomedical-specific instances seem to perform similarly on unseen data.

The results presented in this paper correspond to a single training iteration on all combinations of the pre-defined train-test splits of the data. For future work, we consider performing proper cross-validation experiments by dividing each corpus into folds and using all resulting combinations in order to obtain more stable results and confidence intervals. However, this approach increases the number of experiments from 34 trained models to 340. We have obtained preliminary results of such evaluation, and the findings are consistent with the ones presented in this work. Other options for further research include extending our comparison of multiple Transformers, or even considering generative approaches to NER.

Investigating whether other label alignments are meaningful once more comparable datasets become available would help to reinforce our results outside of medication annotations. Given the currently available corpora, the only other entities that might be comparable are the short version of *Diagnosis/Pathology* and *Therapeutic Procedure* from GGPONC 2.0 with *Diagnosis* and *Treatment* from BRONCO150, respectively. However, here the differences in semantics are even more pronounced than in the case of medication. Annotation campaigns using unpublished corpora, which concern other medical fields and text genres, suggest that we might be able to harmonize other semantic classes in the future. For instance, the *Condition* category in the fine-grained annotation scheme proposed by Roller et al. (2016) for clinical notes in nephrology roughly corresponds to the *Findings* class in GGPONC 2.0.

To conclude, our study calls for more representative large German clinical corpora to generate robust NER taggers that can be used for real-world scenarios, together with a consensus on the semantics and annotation guidelines to equate labeled entities through the datasets.

Limitations

Our findings are limited to medication entities, the only semantic class that is annotated in all available corpora. Moreover, we had to exclude BRONCO150 for long-span experiments due to a mismatch of entity definitions. Although the label alignment decisions are somewhat subjective, they are made based on a thorough inspection of definitions and samples.

The differences in annotation quality and biases may be playing an uncertain role in the models. However, making statements on the impact of the annotation quality is challenging, since each work followed a different annotation protocol and reports different measures of annotator agreement. This is another area where harmonization efforts might be warranted for future research. Furthermore, exploring different hyperparameter configurations lied out of scope for our work, but could have a substantial impact. Mainly, the results from the Transformers comparison (Table 6) could shed different conclusions if the hyperparameters were optimized for each model.

Acknowledgements

Parts of this work were generously supported by grants of the German Federal Ministry of Research and Education (01ZZ1802H, 01ZZ2314N) and the German Federal Ministry of Economic Affairs and Climate Action (01MJ21002A).

References

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Florian Borchert, Christina Lohr, Luise Modersohn, Thomas Langer, Markus Follmann, Jan Philipp Sachs, Udo Hahn, and Matthieu-P. Schapranow. 2020. GGPONC: A corpus of German medical text with rich metadata based on clinical practice guidelines. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages

38–48, Online. Association for Computational Linguistics.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Keno K Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P Loyen, Stefan M Niehues, et al. 2023. MEDBERT.de: A comprehensive German BERT model for the medical domain. *arXiv preprint arXiv:2303.08179*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Donald C Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, et al. 2013. BioC: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Kevin Donnelly. 2006. SNOMED-CT: the advanced terminology and coding system for eHealth. In *Medical and Care Compunetics 3*, number 121 in Studies in Health Technology and Informatics, pages 279–290, Amsterdam etc. IOS Press.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM Sigkdd Explorations Newsletter*, 12(1):49–57.

Johann Frei and Frank Kramer. 2022. GERNERMED: An open German medical NER model. *Software Impacts*, 11:100212.

Johann Frei and Frank Kramer. 2023. German medical named entity recognition model and data set creation using machine translation and word alignment: Algorithm development and validation. *JMIR Form Res*, 7:e39077.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose Posada, John Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. BigBio: A framework for data-centric biomedical natural language processing. In *Advances in Neural Information Processing Systems*, volume 35, pages 25792–25806. Curran Associates, Inc.

Dieter Galea, Ivan Laponogov, and Kirill Veselkov. 2018. Exploiting and assessing multi-source data for supervised biomedical named entity recognition. *Bioinformatics*, 34(14):2474–2482.

John M Giorgi and Gary D Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Udo Hahn, Franz Matthies, Christina Lohr, and Markus Löffler. 2018. 3000pa-towards a national reference corpus of german clinical language. In *MIE*, pages 26–30.

Suwisa Kaewphan, Sofie Van Landeghem, Tomoko Ohta, Yves Van de Peer, Filip Ginter, and Sampo Pyysalo. 2016. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025.

Maximilian König, André Sander, Ilja Demuth, Daniel Diekmann, and Elisabeth Steinhagen-Thiessen. 2019. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. *PloS one*, 14(11):e0224916.

Manuel Lentzen, Sumit Madan, Vanessa Lage-Rupprecht, Lisa Kühnel, Juliane Fluck, Marc Jacobs, Mirja Mittermaier, Martin Witzenrath, Peter Brunecker, Martin Hofmann-Apitius, et al. 2022. Critical assessment of transformer-based ai models for german clinical notes. *JAMIA open*, 5(4):ooac087.

Ignacio Llorca. 2022. Programmatic access to FairEval as a HuggingFace evaluation module. `https://huggingface.co/spaces/hpi-dhc/FairEval`. (Last accessed: April 26th, 2023).

Ignacio Llorca. 2023. BIGBIO loaders for German clinical corpora (GGPONC 2.0, CARDIO:DE, BRONCO150) in the HuggingFace Hub. `https://huggingface.co/datasets/bigbio/`{ggponc2, cardiode, bronco}. (Last accessed: April 26th, 2023).

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution — a case study to get around

IPRs and privacy constraints featuring the German JSYNCC corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO - the first publicly shareable, multiply-alienated german clinical text corpus. *Studies in health technology and informatics*, 296:66—72.

Hiroki Nakayama. 2018. seqeval: A Python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Joel Nothman, Tara Murphy, and James R. Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 612–620, Athens, Greece. Association for Computational Linguistics.

Katrin Ortmann. 2022. Fine-grained error analysis and fair evaluation of labeled spans. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1400–1407, Marseille, France. European Language Resources Association.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Mingyang He, Michael M. Allers, Anna S. Tiefenbacher, Nicola Kunz, Anna Martynova, Noemie Spiller, Julian Mierisch, Florian Borchert, Charlotte Schwind, Norbert Frey, Christoph Dieterich, and Nicolas A. Geis. 2023. A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Nature Scientific Data*, 10:207.

Roland Roller, Laura Seiffe, Ammer Ayach, Sebastian Möller, Oliver Marten, Michael Mikhailov, Christoph Alt, Danilo Schmidt, Fabian Halleck, Marcel Naik, et al. 2022. A medical information extraction workbench to process german clinical text. *arXiv preprint arXiv:2207.03885*.

Roland Roller, Hans Uszkoreit, Feiyu Xu, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. 2016. A fine-grained corpus annotation schema of German nephrology records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 69–77, Osaka, Japan. The COLING 2016 Organizing Committee.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. GottBERT: a pure German language model. *arXiv preprint arXiv:2012.02110*.

Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. 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*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Torsten Zesch and Jeanette Bewersdorff. 2022. German medical natural language processing–a data-centric survey. In *The Upper-Rhine Artificial Intelligence Symposium UR-AI 2022 : AI Applications in Medicine and Manufacturing, 19 October 2022, Villingen-Schwenningen, Germany*, pages 137–145. Furtwangen University.

Zixing Zhang, Felix Weninger, Martin Wöllmer, and Björn Schuller. 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 523–528.