Drug Safety (2023) 46:765–779
https://doi.org/10.1007/s40264-023-01322-3

ORIGINAL RESEARCH ARTICLE

Provision and Characterization of a Corpus for Pharmaceutical, Biomedical Named Entity Recognition for Pharmacovigilance: Evaluation of Language Registers and Training Data Sufficiency

Jürgen Dietrich[1] · **Philipp Kazzer**[2]

Accepted: 16 May 2023 / Published online: 20 June 2023

Abstract

Introduction and Objective Machine learning (ML) systems are widely used for automatic entity recognition in pharmacovigilance. Publicly available datasets do not allow the use of annotated entities independently, focusing on small entity subsets or on single language registers (informal or scientific language). The objective of the current study was to create a dataset that enables independent usage of entities, explores the performance of predictive ML models on different registers, and introduces a method to investigate entity cut-off performance.

Methods A dataset has been created combining different registers with 18 different entities. We applied this dataset to compare the performance of integrated models with models created with single language registers only. We introduced fractional stratified k-fold cross-validation to determine model performance on entity level by using training dataset fractions. We investigated the course of entity performance with fractions of training datasets and evaluated entity peak and cut-off performance.

Results The dataset combines 1400 records (scientific language: 790; informal language: 610) with 2622 sentences and 9989 entity occurrences and combines data from external (801 records) and internal sources (599 records). We demonstrated that single language register models underperform compared to integrated models trained with multiple language registers.

Conclusions A manually annotated dataset with a variety of different pharmaceutical and biomedical entities was created and is made available to the research community. Our results show that models that combine different registers provide better maintainability, have higher robustness, and have similar or higher performance. Fractional stratified k-fold cross-validation allows the evaluation of training data sufficiency on the entity level.

Key Points

Combined language register models provide similar results or outperform single language register models.

A new method enables the evaluation of training data sufficiency for computer system validation.

1 Introduction

The efficient and effective detection of adverse events (AEs) in free text (e.g., screening of literature, spontaneous, study, and regulatory reports) is required by Good pharmacovigilance practices (GVP) in the European Union (EU) (EU-GVP) Modules and is of paramount importance for pharmacovigilance. Machine learning (ML)-based models have been developed to support this task and to help get efficiencies in this process, especially as content that needs to be screened has increased over the last few years. Several initiatives are ongoing to investigate the industry perspective on artificial intelligence (AI) in pharmacovigilance [1–4] to evaluate risks and to improve scalability, efficiency, and quality. There is a growing industrial interest in using cognitive technologies, especially in case intake and case processing.

✉ Jürgen Dietrich
juergen.dietrich@bayer.com

[1] Bayer AG, Pharmaceuticals, Medical Affairs & Pharmacovigilance, Data Science & Insights, Müllerstr. 170, 13353 Berlin, Germany

[2] Syncwork AG, Systems Development, Berlin, Germany

Transformer-based bidirectional encoder language models (BERT [5]) and their variants (e.g., ClinicalBERT [6], BioBERT [7]) are widely used in AE detection (e.g., [8]) and carry out predictions with a decoder implemented as a task-specific trainable network. Raffel et al. [9] introduced a Text-To-Text Transfer Transformer (T5), a pretrained encoder-decoder transformer.

These models are further fine-tuned with data for specific pharmacovigilance tasks; however, there are typical challenges with small, variable, and imbalanced datasets and terms used very differently and often vaguely or inaccurately. In previous studies, the T5 architecture demonstrated high flexibility in dealing with text from different domains and ontologies [10, 11]. We used in our experiments T5 since it was shown that T5 outperformed other AE detection models based on BERT variants and can be used to apply the same model for multiple and diverse datasets [12].

Training data are required to train ML models, and those models can only be as good as the training data used to train them. Publicly available pharmacovigilance datasets (e.g., ADE Corpus V2 [13], WEB-RADR [14], CADEC [15], SMM4H [8, 16], BioCreative VII Task 3 [17]) focus on small subsets of annotated entities (e.g., drug, dosage, AE, indication), and single language registers (social media, scientific literature) contain different entity subsets or do not follow an annotation scheme that allows the use of annotated entities independently from other entities. In this paper, we describe our activities to generate, characterize, and consistently re-annotate a dataset collection by completing annotations on pharmaceutical and biomedical entities over all sources. We annotated besides AE other medical entities (e.g., indication, comorbidity) in order to enable fine-tuned ML systems to distinguish between identical medical terms context-wise.

ML models trained on limited entity subsets or covering one single register do not take full advantage of the available annotations and become specialized for a given text type, requiring potentially additional efforts to be maintained. However, having the maintainability, robustness, and performance of a production system in mind, several challenges arise. It is plausible that the effort to maintain multiple systems increases compared to a single integrated system (a model trained on multiple language registers). In our dataset collection, we intentionally combined datasets from different language registers and fine-tuned and tested models using k-fold cross-validation [18, 19], stratified on entities and data sources, to achieve a similar distribution of target class labels and language registers. We used ML models trained on stratified data to investigate in our first experiment whether the performance of an integrated system is comparable to models trained on single language registers only.

Currently, there is an ongoing discussion on how computer system validation (CSV) can be applied in scenarios where ML and AI applications are used in regulated environments. The Council for International Organizations of Medical Sciences (CIOMS) Working Group XIV on Artificial Intelligence in Pharmacovigilance was established to provide recommendations on validation, machine training, and evaluation of the appropriateness of the different tools, and how they can be implemented and maintained. The aspect regarding the amount of training data needed is considered as crucial [2].

Past studies showed that different pharmaceutical entities are harder to predict than others by a given model [20]. It is obvious that the semantic complexity of an entity influences the prediction performance and the amount of training data necessary to be provided to reach peak performance. In a second experiment, we introduced fractional stratified k-fold cross-validation to evaluate the entity cut-off performance by increasing training sample sizes to determine training data sufficiency in CSV tasks.

2 Methods

2.1 Preparation of the Dataset

2.1.1 Data Sources

Records were randomly selected to integrate different language registers from external (ADE Corpus V2 [based on PubMed]: 500 [13]; SMM4H [based on Twitter]: 100 [16]; drugs.com: 201) and Bayer internal data sources (Bayer Literature database: 89; Safety and Product Technical Complaint databases: 510 records). There are 1400 records in total, consisting of one or multiple sentences. Data from external and internal sources were retrieved between March and September 2021.

ADE Corpus V2: This dataset [13] contains case reports extracted from MEDLINE, with annotations of drug, dosage, and AE. From ADE Corpus, we randomly selected 400 *positive* records (i.e., containing at least one drug-related AE mention) and 100 *negative* records (i.e., do not contain information about adverse drug effects [ADEs]).

SMM4H: This dataset was introduced for the Shared Tasks on AE in the workshop on Social Media Mining for Health Applications (SMM4H) [16]. The dataset is composed of Twitter posts, typically short, informal texts with non-standard orthography, and it contains annotations for both detection (i.e., task 1, binary classification) and extraction (i.e., task 2, named entity recognition) of AEs. We randomly selected 100 records from SMM4H 2017 and 2019 datasets.

Drugs.com: Records were retrieved on 30 September 2021 from the Food and Drug Administration (FDA) approval history section [21] between 1 January 2010 and 30 September 2021 by means of a robotic process automation tool (UiPath). After retrieval, about 200 records were randomly selected for annotation.

Bayer Literature database: 53 PubMed abstracts of clinical study results were retrieved from the Bayer internal literature database, and were split into sentences; 89 records were randomly sampled.

Safety and Product Technical Complaint databases: This dataset contains data from two sources (Bayer Safety database: 260 records; Product Technical Complaint (PTC) database: 250 records), consisting of one or multiple sentences per record. Both datasets combine mainly patient or consumer recorded reports retrieved from call center records, emails, etc., mainly in informal language, combining information about product quality issues and AEs, e.g., *'I'm having like heavy bleeding like 21 days or something and really sharp pain.'*

2.1.2 Entities

In total, 18 different biomedical and pharmaceutical entities (e.g., drug, dose, AE, indication, comorbidity, intended use, administration route) were manually annotated and complemented in all datasets, and if required, existing annotations were corrected. The selection of the entities aimed to categorize a given text from a pharmacovigilance perspective as completely as possible. Texts from various sources contain potentially different entities.

The dataset comprises, in total, 1400 records, with 2622 sentences, 9989 entity occurrences (for details, see Tables 1, 2), and 6185 occurrences of one or more entities on the record level. The number of records with scientific language is 790; the number with informal ('lay person') language is 610. The set combines data from external (801 records) and internal sources (599 records); see Table 2 for more details.

2.1.3 Data Quality

Data preparation The ADE Corpus and SMM4H datasets were converted to single lines for each PubMed ID or Tweet ID, respectively. Multiple entity occurrences in one record (e.g., AE, drug) were aggregated into one field separated with semicolons, and Tweets were finally normalized (e.g., substitution of user nicknames with @USER).

Quality measures The quality of the annotations was ensured by annotation guidelines, team agreements, measuring inter-annotator agreement (IAA) against ADE Corpus and SMM4H dataset by determination of Cohen's kappa and Gwet AC1, multi-stage annotations, and final review. Datasets were reviewed and annotated in total by five experienced pharmacovigilance case processors. This task was executed independently of the original annotation of publicly available datasets. In periodic team meetings, newly annotated records were reviewed and, if required, corrected by team agreement. Internal data retrieved from our Safety and Product Technical Complaint databases underwent internal quality control measures (results are not published here).

Use of T5 model to improve data quality We used a 'machine-in-the-loop' approach to improve the quality and consistency of the dataset. During multi-stage annotation cycles, we used T5 models trained on the actual dataset and reviewed manually the false-positive and false-negative results from fivefold cross-validation runs to detect through the complete dataset the incompleteness and inconsistencies of our previous annotations.

Inter-annotator Agreements IAA methods are described in the electronic supplementary material (ESM 3, Sect. 1.1).

Annotation guidelines Guidelines were adapted from CADEC annotation guidelines [15] and provided to all annotators, defining the rules that annotators should follow when working on documents.

The following adaptations were used:

- Annotations were done at the paragraph level, i.e., we annotated the complete record.
- The entity description (see Table 1) was used for annotator training and final review.
- Entities were annotated independently of the occurrence of other entities.
- All entity mentions were annotated in the sequence of occurrence separated by semicolons. Duplicate entities within one record were annotated independently; that means, all occurrences of the same entity were annotated.
- The identification of indication (the reason for using a drug) versus comorbidity versus medical history/condition were executed on the basis of plausibility (e.g., *'We report a 31-year-old woman with recurrent Hodgkin's lymphoma and unrecognized HMSN-1 who developed severe motor neuropathy 3 weeks after the first cycle of treatment including 2 mg of vincristine'*; indication: recurrent Hodgkin's lymphoma; comorbidity: hereditary motor and sensory neuropathy type 1 [HMSN-1]).
- Abbreviations of drugs (e.g., MTX for methotrexate), methods (e.g., HAIC for hepatic arterial infusion chemotherapy), indications (e.g., HFS for hand-foot syndrome), etc. were annotated. Metabolic products (e.g., *desmethylsertraline blood levels*) and blood parameters (e.g., *increased serum lithium concentrations*) were excluded from annotation.
- Hashtags (e.g., *#crohns* or *#ExtremePain*) in Twitter data (SMM4H) were evaluated and annotated if required.

Table 1 Entities and their description and the number of entities annotated (#Entity Occurrences)

Entity	Description	#Enitiy occurrences	In %
Action	Introduced to describe the mode of action (e.g., *multikinase inhibitor*, *TKI*), we used it also to categorize a drug or methodology in more detail (e.g., *gonadotropin releasing hormone [GnRH] analog*) or to give additional information about substance class (e.g., *broad-spectrum antibiotic*)	213	2.13%
Administration form/primary packaging	Used to describe the administration form of a medication (e.g., *gel*, *tablet*, *ointment*, *contrast agent*), give some information about the form (e.g., *liquid*) or the primary or secondary packaging (e.g., *filler*, *bottle*, *syringe*)	491	4.92%
Administration route	Used to describe the administration route of a medication (e.g., *p.o.*, *intravenous*)	164	1.64%
AE	Any occurrence of a side effect for a patient that could be potentially caused by a drug or medical device (this may include also drug ineffectiveness, death, unexpected beneficial drug effects, or unexpected therapeutic responses), in the case of medical devices, any PTC occurrence in which the device could had direct contact with a patient/consumer is also reported as AE	1602	16.04%
Comorbidity	Refers to further conditions (concurrent conditions or co-infections) that define the patient population apart from the 'disease/symptoms/procedure' aspect of the indication (e.g., *treatment of pancreatic insufficiency in patients with cystic fibrosis*; indication: pancreatic insufficiency; comorbidity: cystic fibrosis). Due to sparse data in the original text, there are potential problems regarding differentiating comorbidity from indication or medical history/condition	143	1.43%
Dose	Any specific quantitative measurements (e.g., *0.3 mg/kg/day*, *50 Gy*), frequency mentions (e.g., *two tablets twice a day*), or unspecific mentions (e.g., *overdose*, *high dose*) that describe the dosage of drug or methodology	343	3.43%
Drug/device	Used to specify a drug or device by, e.g., INN, trade name, IUPAC name, abbreviations, e.g., *MTX*. This category does not include the names of metabolites or reaction byproducts	1504	15.06%
Indication	Used to specify the reason for using a drug, device, or methodology. This term is used intentionally not in a regulatory sense of a registered medication, in order to detect potential off-label use (e.g., *cutaneous T cell lymphoma* includes abbreviations *CTCL* and *ADHD;* informal terms: *drippy nose*, *for my heart*)	787	7.88%
Intended effect	Used to specify the result/type of outcome intended for the target condition, aim, or strategy to be achieved by the indication (e.g., *treatment*, *prophylaxis*, *replacement therapy*, *chemotherapy*, *prevention*)	448	4.48%
Medical history/condition	Used to describe past condition or medical occurrence. In cases where it is not clear whether the medical occurrence persists, sicknesses are annotated as comorbidities	50	0.50%
Method/procedure/administration	Used to describe a procedure (e.g., *colonoscopy prep*, *insertion*, *removal*, *dilate*) or methodology (e.g., *colonoscopy*, *ultrasound*, *MRI*, *HAIC*)	402	4.02%
Outcome	Mainly used to describe the reached outcome of a study or activity (e.g., *disease stabilization*, *demonstrated antitumor activity and safety*)	42	0.42%
PDC	PDC specifies whether dose information refers to a specific drug; the syntax is <Product>|<Dose>. In the case of uncertainty, all products mentioned are annotated	337	3.37%
PEC	PEC specifies whether an AE refers to a drug; the syntax is <Product>|<Event>. In the case of uncertainty, all products mentioned are annotated	1420	14.22%
PIC	PIC specifies whether an indication refers to a drug; the syntax is <Product>|<Indication>. In the case of uncertainty, all products mentioned are annotated	1082	10.83%
PTC	Any occurrence that may affect the quality of a product that is caused by the manufacturing process (including delivery, storage, counterfeit) or the design of a medication or device (e.g., *4 tablets were missing*, *the pill has a different color*, *a piece of a 2–3 mm piece of IUD missing*, *IUD was expelled*, or *removal of IUD*)	693	6.94%
Target parameter	Used to characterize study or experiment targets (e.g., *efficacy and toxicity*, *prolong survival*, *6 months PFS of 50% or greater*)	31	0.31%

Table 1 (continued)

Entity	Description	#Enitiy occurrences	In %
Target population	Used to describe—specifically or unspecifically—a subpopulation of an experiment (*smaller premature infants*), study (*post-menopausal women*), or single case report (e.g., *7-year-old child*). It contains mainly demographic information, but could also list inclusion criteria of a study or a condition for an outcome of an investigation. Mentions of sicknesses are excluded and annotated as comorbidity, indication, or medical history/condition	237	2.37%
	Sum	9989	100%

ADHD attention deficit hyperactivity disorder, *AE* adverse event, *CTCL* cutaneous T cell lymphoma, *HAIC* hepatic arterial infusion chemotherapy, *INN* international nonproprietary name, *IUD* intrauterine device, *IUPAC* International Union of Pure and Applied Chemistry, *MRI* magnetic resonance imaging, *MTX* methotrexate, *PDC* product dose combination, *PEC* product event combination, *PFS* progression-free survival, *PIC* product indication combination, *p.o.* per os (oral administration), *PTC* product technical complaint, *TKI* tyrosine kinase inhibitor

Table 2 Overall distribution of records over the various entities and sources

	Safety and product technical complaint databases[a]	ADE Corpus V2[b]	drugs.com[b]	Bayer Literature database[b]	SMM4H[a]	Sum (occurrence on record level)
Action	1	11	174	9	3	198
Administration form/primary packaging	232	15	40	0	7	294
Administration route	17	70	56	5	2	150
AE	243	454	2	33	76	808
Comorbidity	2	56	40	10	2	110
Dose	2	252	25	9	6	294
Drug/device	172	452	201	89	93	1007
Indication	91	185	201	66	18	561
Intended effect	9	151	192	45	4	401
Medical history/condition	11	30	1	0	0	42
Method/procedure/administration	132	95	39	17	4	287
Outcome	3	11	0	19	0	33
PDC	0	239	3	7	5	254
PEC	120	428	2	32	76	658
PIC	17	164	201	66	15	463
PTC	382	1	0	0	0	383
Target parameter	0	3	0	20	0	23
Target population	13	125	60	18	3	219
Sum (occurrence on record level)	1447	2742	1237	445	314	6185
Record #	510	500	201	89	100	1400

AE adverse event, *ADE* adverse drug effect, *PDC* product dose combination, *PEC* product event combination, *PIC* product indication combination, *PTC* product technical complaint

[a]Safety and Product Technical Complaint databases, SMM4H: sources that contain mainly informal language

[b]ADE Corpus V2, drugs.com, Bayer Literature database: sources with scientific language. Note that only the occurrence of one or more entities on record level is shown

- Product combinations (Product Dose Combination [PDC], Product Event Combination [PEC], Product Indication Combination [PIC]) were annotated only if the product was mentioned and related (e.g., *'Central nervous system manifestations of an ibuprofen overdose reversed by naloxone'*; PDC: *ibuprofen/overdose;* PEC: *ibuprofen/overdose;ibuprofen/central nervous system manifestations;* PIC: *naloxone/central nervous system manifestations of an ibuprofen overdose*). In the case of

uncertainty between PEC and PIC, both product combinations were annotated.
- Annotations were executed per record (i.e., no external information was used). However, for the identification, e.g., of drug–drug interactions or overdose, additional sources were used (e.g., drugs.com or Summary of Product Characteristics [SmPCs]).

Terms were extracted as completely as possible to facilitate medical judgement (e.g., *secondary cardiac complications* instead of *cardiac complications* or *subacute encephalopathy* instead of *encephalopathy* or *psychotic reaction disappeared* instead of *psychotic reaction*).

- According to EU-GVP Module VI, any occurrence of death should be medically evaluated, and therefore, from a risk-based approach, we annotate 'death' as an AE to enable detection by ML systems and to ensure medical judgement.

2.2 T5 Modeling

2.2.1 Pretraining

There are various sizes available for pretrained T5 models [22] based on Colossal Clean Crawled Corpus (C4), including small (60 million parameters), base (220 million parameters), large (770 million parameters), 3B (3 billion parameters), and 11B (11 billion parameters). In the paper, we use the term 'T5' to refer to the architecture 'T5-base.' We used this version because it provides a good trade-off between speed and accuracy.

2.2.2 Data Preparation for Model Training

Datasets used for our experiments underwent an 80:20 split on the record level, if not otherwise specified. Models were trained on 80% of the data and validated using the remaining 20% of the data by comparing the model predictions with the actual annotations made by the annotation team. In the case of fivefold cross-validation, in total, five subsequent non-overlapping 20% validation datasets were selected to validate the complete dataset and the remaining 80% of data was used as a training dataset. A stratification on source and entity was applied on validation and training data. Finally, records available within each training dataset were randomly sampled before model fine-tuning.

2.2.3 Model Fine-Tuning

In our experiments, we used an Adam optimizer with weight decay and set the maximum sequence length to 256. The learning rate was set to 1*e*-4, batch size was set to 4, and the epoch was set to 3. We performed a greedy search.

2.2.4 Evaluation

The evaluation was executed per entity type and based on per positive class values of F score metrics. A term is correctly detected only if the system is able to assign the correct prediction label and correct entity type according to the International Workshop on Semantic Evaluation [23]. A correct prediction with incorrect type is considered as 'missing' (false negative) and a correct type with an incorrect prediction as 'spurious' (false positive).

Please note the definition of 'strict' and 'partial' matches:

- Strict: exact boundary surface string match (i.e., spans of prediction and truth are identical).
- Partial: partial boundary match over the surface string (i.e., spans of prediction and truth are overlapping).

For multiclass evaluations, the macro-weighted averaged F1 scores were calculated (weighted by class frequency and on positive class) [24].

2.3 Setup of the Experiments

2.3.1 Experiment 1: Integrated Model Versus Single Language Register Trained Models

In this experiment, we investigated the effect of pure single language register systems versus an integrated system. We set up stratified fivefold cross-validation runs [25]. We were starting with models fine-tuned on 100% informal ('Lay')

Table 3 Number of informal ('lay person' [Lay]) and scientific (Sc) records used for investigation of integrated versus single language register models

	Proportion of informal: scientific training data					
	Lay100Sc0 (100%:0%)	Lay80Sc20 (80%:20%)	Lay60Sc40 (60%:40%)	Lay40Sc60 (40%:60%)	Lay20Sc80 (20%:80%)	Lay0Sc100 (0%:100%)
Informal records	486	388	291	194	97	0
Science records	0	98	195	292	389	486

The percentages of training data coming from Lay and Sc corpus are shown in parentheses

data and subsequently substituting 20% of the informal data by the same amount of scientific (Sc) data until 100% science data was reached (see Table 3). Note that we kept the amount of training data the same and focused on the effect of the multiple language domains.

Three analyses were performed on all models fine-tuned with different proportions of language registers. Within each analysis, the model performance was tested with the same validation dataset.

In our first analysis, we used AE and indication, since both entities were available and showed a high variability in different language registers. The model performance was tested with an identical validation dataset consisting of 50% informal and 50% scientific data (124 records each). We mainly focused on the pure language register models (Lay100Sc0 and Lay0Sc100). We were expecting an inverted U-shaped curve, but wanted to evaluate how pronounced the performance of the pure language register models decreased.

In the other two analyses, we focused on indication and tested the model performance with (1) 100% of the informal dataset and (2) 100% of the scientific dataset to determine the performance decrease of models trained with proportions of the opposite language register.

2.3.2 Experiment 2: Fractional Stratified Fivefold Cross-Validation

To investigate the development of the entity performance, we used in our second experiment a newly developed method (fractional stratified k-fold cross-validation) to fine-tune models with proportions of training data. In contrast to the previously described experiment in Sect. 2.3.1, we ensured that all validation and training data were stratified on entities and sources equally. In this experiment, we used a modification of the stratified fivefold cross-validation approach with an 80:20 split. We split the complete dataset into five stratified 20% folds and selected for each run a 20% fold as a validation dataset and used the remaining 20% folds for the creation of 20%, 40%, 60%, and 80% training sets, i.e., one cross-validation fold contains about 20% of entity occurrences (see Table 1). Per 20% validation dataset, four 20%, two 40%, two 60%, and one 80% training folds were created. We repeated this procedure five times to use all data for validation ($n = 45$). We ensured that for each 20%-fold validation dataset, all training data were used for model fine-tuning, but all permutations were used only for the 20% (20 models) and 80% folds (five models), due to the effort involved. For 40% folds, two adjoined folds were selected for training, and for 60% folds, three adjoined folds were selected for training (e.g., validation dataset: 20% fold #1; first 40%-fold training set: 20% fold #2 and #3; second 40% fold: #4 and #5; first 60% fold: #2–#4; second 60% fold: #3–#5). With all individual training sets, T5 models were fine-tuned and evaluated. In one experiment, we decreased the validation and training folds to 10% and determined the 10% data point for AE (additional nine 10% folds in ten runs, $n = 90$). Please note that our decision to use 20% folds for this experiment is based on fivefold cross-validation, but is in principle arbitrary. This choice is from our perspective a good compromise between the evaluation of performance details and effort spent on the model creation.

The methods of the binary classifier experiment are described in ESM 3, Sect. 2.1 (see the electronic supplementary material).

3 Results

3.1 Dataset

The dataset provided as an Excel spreadsheet (see ESM 1 in the electronic supplementary material) combines the annotated entities per record in columns from the following sources (in total, 1400 records; only 890 records selected for publication):

- ADE Corpus V2 (500 records).
- drugs.com (201 records).
- SMM4H (100 records).
- Bayer Literature database (89 records).

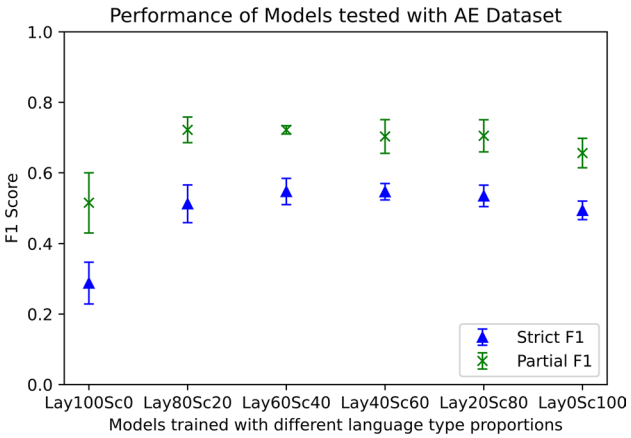

Fig. 1 AE strict and partial F1 scores (mean and standard deviation) for models fine-tuned with different portions of informal and scientific data; test dataset: 50% informal and 50% scientific data. Proportion of informal to scientific data: *Lay100Sc0* 100%:0%, *Lay80Sc20* 80%:20% … *Lay0Sc100* 0%:100%. *AE* adverse event, *Lay* lay person/informal, *Sc* scientific

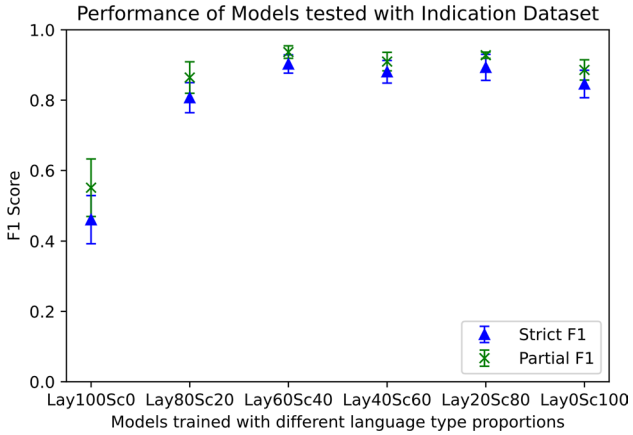

Fig. 2 Indication strict and partial F1 scores (mean and standard deviation) for models fine-tuned with different portions of informal and scientific data; test dataset: 50% informal and 50% scientific data. Proportion of informal to scientific data: *Lay100Sc0* 100%:0%, *Lay80Sc20* 80%:20% … *Lay0Sc100* 0%:100%. *Lay* lay person/informal, *Sc* scientific

- Bayer Safety and Product Technical Complaint databases (510 records, not published).

The spreadsheet lists the unique ID of the record, the source, PubMed ID (if available), input_text, and all entities listed in Table 1 in separate columns. In addition, we enclosed a spreadsheet that documents the spans of each annotation in the original record (*input_text*) (see ESM 2).

3.2 Experiment 1: Integrated Model Versus Single Entity and Single Language Register Trained Models

As described in Sect. 2.3.1, we investigated the performance of models fine-tuned with different proportions of informal ('Lay') and scientific data stratified for two entities (AE and indication) in fivefold cross-validation runs. Results are shown in Figs. 1 and 2 and summarized in Table 4.

We observed that even a small portion of 20% scientific data and 80% informal data increased the performance for

Table 4 Results from models fine-tuned with different portions of scientific and informal data and different test datasets and target entities (AE, indication)

No.	Model	Test dataset	Entity	Figure	F1 type	F1 mean ($n = 5$)	Standard deviation
1	Lay100Sc0	Sc: 50%/Inf: 50%	AE	4	Strict	0.288	0.059
2	Lay100Sc0	Sc: 50%/Inf: 50%	AE	4	Partial	0.515	0.085
3	Lay0Sc100	Sc: 50%/Inf: 50%	AE	4	Strict	0.493	0.026
4	Lay0Sc100	Sc: 50%/Inf: 50%	AE	4	Partial	0.656	0.042
5	Lay60Sc40	Sc: 50%/Inf: 50%	AE	4	Strict	0.547	0.037
6	Lay60Sc40	Sc: 50%/Inf: 50%	AE	4	Partial	0.722	0.012
7	Lay100Sc0	Sc: 50%/Inf: 50%	Indication	5	Strict	0.461	0.068
8	Lay100Sc0	Sc: 50%/Inf: 50%	Indication	5	Partial	0.551	0.082
9	Lay0Sc100	Sc: 50%/Inf: 50%	Indication	5	Strict	0.846	0.039
10	Lay0Sc100	Sc: 50%/Inf: 50%	Indication	5	Partial	0.886	0.029
11	Lay60Sc40	Sc: 50%/Inf: 50%	Indication	5	Strict	0.903	0.026
12	Lay60Sc40	Sc: 50%/Inf: 50%	Indication	5	Partial	0.937	0.018
13	Lay100Sc0	Sc: 100%	Indication	6	Strict	0.424	0.087
14	Lay100Sc0	Sc: 100%	Indication	6	Partial	0.479	0.085
15	Lay0Sc100	Sc: 100%	Indication	6	Strict	0.936	0.028
16	Lay0Sc100	Sc: 100%	Indication	6	Partial	0.938	0.027
17	Lay60Sc40	Sc: 100%	Indication	6	Strict	0.958	0.011
18	Lay60Sc40	Sc: 100%	Indication	6	Partial	0.959	0.010
19	Lay100Sc0	Inf: 100%	Indication	7	Strict	0.783	0.098
20	Lay100Sc0	Inf: 100%	Indication	7	Partial	0.797	0.093
21	Lay0Sc100	Inf: 100%	Indication	7	Strict	0.574	0.117
22	Lay0Sc100	Inf: 100%	Indication	7	Partial	0.598	0.107
23	Lay60Sc40	Inf: 100%	Indication	7	Strict	0.811	0.076
24	Lay60Sc40	Inf: 100%	Indication	7	Partial	0.826	0.063

Model: Proportion of Inf to Sc data: *Lay100Sc0* 100%:0%, *Lay60Sc40* 60%:40%, *Lay0Sc100* 0%:100%. Test dataset: proportion of Sc data; proportion of Inf data

AE adverse event, *Inf* informal, *Lay* lay person/informal, *Sc* scientific

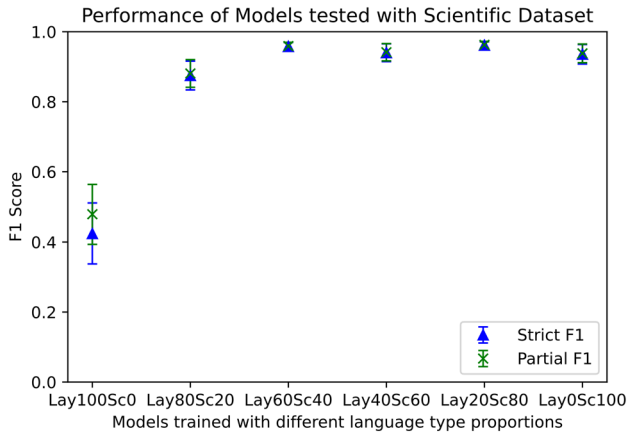

Fig. 3 Indication strict and partial F1 scores (mean and standard deviation) for models fine-tuned with different portions of informal and scientific data; test dataset: 100% indication scientific data. Proportion of informal to scientific data: *Lay100Sc0* 100%:0%, *Lay80Sc20* 80%:20% … *Lay0Sc100* 0%:100%. *Lay* lay person/informal, *Sc* scientific

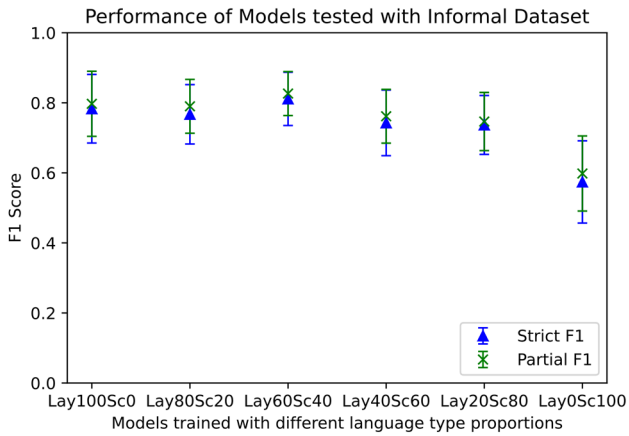

Fig. 4 Indication strict and partial F1 scores (mean and standard deviation) for different portions of informal and scientific data; test dataset: 100% indication informal data. Proportion of informal to scientific data: *Lay100Sc0* 100%:0%, *Lay80Sc20* 80%:20% … *Lay0Sc100* 0%:100%. *Lay* lay person/informal, *Sc* scientific

AE and indication detection. Models trained with pure scientific data performed better than models trained with pure informal data (see Table 4; AE: rows 1, 2 vs 3, 4; indication: rows 7, 8 vs 9, 10). Both single language register models underperformed compared to an integrated model, e.g., Lay60Sc40 (see Table 4; AE: rows 5, 6; indication: rows 11, 12).

In the next analyses, we investigated the influence of different test datasets on our language register models. We used the entity type 'indication' of pure scientific and informal language registers for testing (see Figs. 3, 4). Regarding the scientific dataset in Fig. 3, the curve progression resembles the one shown in Fig. 2. A 20% substitution of informal data with scientific data again increased the performance significantly. A model trained on data with a single language register performed better on the test data from the same language register than on test data of the other language register (Figs. 3, 4, and Table 4 rows 15, 16 vs 21, 22 and rows 19, 20 vs 13, 14). The integrated model Lay60Sc40 shows higher F1 scores compared to the Lay100Sc0 model (Table 4, rows 23, 24 vs 19, 20) tested on pure informal dataset, and slightly higher F1 scores compared to the Lay0Sc100 model tested on pure scientific dataset (Table 4, rows 17, 18 vs 15, 16).

3.3 Experiment 2: Fractional Stratified Fivefold Cross-Validation

In this experiment, we investigated (1) the entity peak and cut-off performance and (2) whether data are sufficient for entity prediction.

In Table 5, the results of all entities including median partial F1 values for the 20% and 80% training folds are listed. For some entities, the average 20%-fold size is small, e.g., for action and dose (43 and 69 occurrences), but results in high F1 scores (0.808 and 0.884). The entity 'dose' seems to reach peak performance level at 20% training data used. Comorbidity and PTC show significant higher 80%-fold F1 scores compared to 20% folds, which indicates that a higher peak level could be probably achieved with more data provided (see Table 5 and also the F1 curve progression in Fig. 5). Data provided for medical history/condition are not sufficient at all. Comparing the results from AE with indication, indication achieved a higher F1 score (see Table 5; 20% fold: AE 0.620, indication 0.796) even with a lower average 20%-fold size (see Table 5; AE: 320; indication: 157). The combination of product and indication (PIC) seems to be more difficult to predict compared to PDC or PEC (see Table 5).

Figure 6 shows the partial F1 for drug, AE, and PEC trained with fractions of the training dataset. The average numbers of all entity occurrences in the 20% fold are similar. Since the AE partial F1 peak performance was almost reached at 20% training data used (see Table 5), the validation and training folds were decreased to 10% (fold size about 160 occurrences) and the partial F1 at 10% determined (median 0.560).

Table 5 Results from fractional stratified fivefold cross-validation run per entity

Entity	Avg 20%-fold size	Median partial F1 20%	Median partial F1 80%
Action	43	0.808	0.861
Administration form/primary packaging	98	0.639	0.682
Administration route	33	0.602	0.684
AE	320	0.620	0.638
Comorbidity	29	0.129	0.439
Dose	69	0.884	0.890
Drug/device	301	0.908	0.937
Indication	157	0.796	0.814
Intended effect	90	0.820	0.844
Medical history/condition	10	–	0.133
Method/procedure/administration	80	0.500	0.536
Outcome	8	0.250	0.243
PDC	67	0.667	0.731
PEC	284	0.495	0.557
PIC	216	0.209	0.370
PTC	139	0.348	0.649
Target parameter	6	0.400	0.311
Target population	47	0.741	0.764

Median partial F1 values for 20% and 80% of training data are listed

AE adverse event, *Avg* average, *PDC* Product Dose Combination, *PEC* Product Event Combination, *PIC* Product Indication Combination, *PTC* Product Technical Complaint

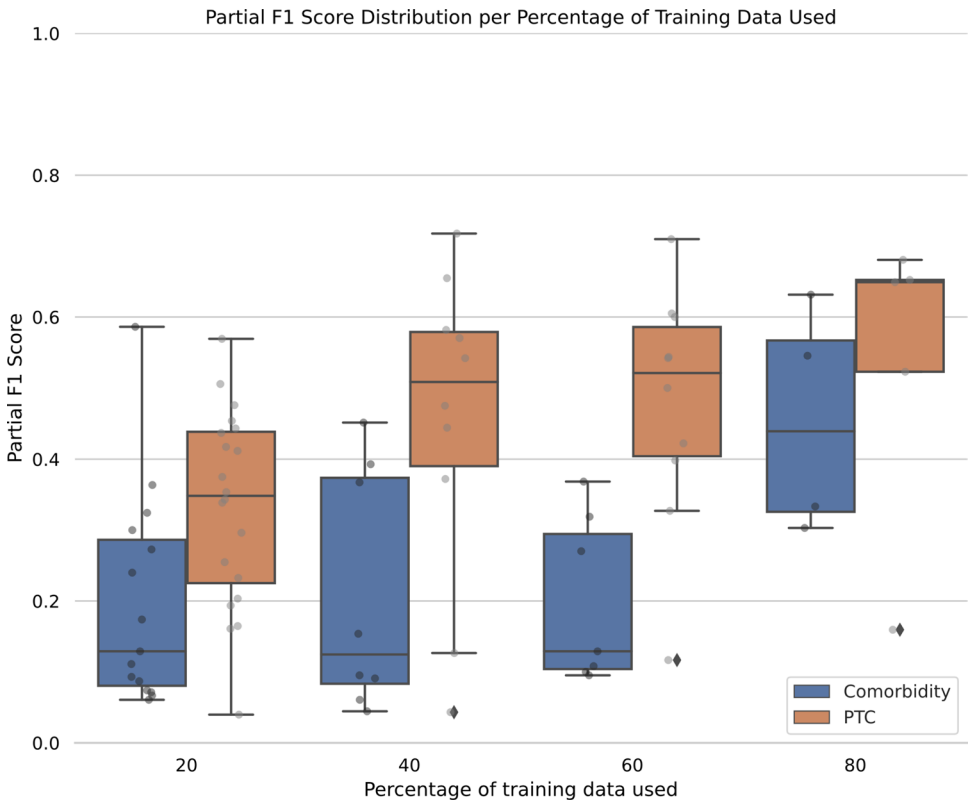

Fig. 5 Boxplots of partial F1 scores of different comorbidity and PTC training data fractions in stratified fivefold cross-validation runs (each 20% fold consists of around 29 [comorbidity] and 139 [PTC] occurrences, respectively). *PTC* Product Technical Complaint

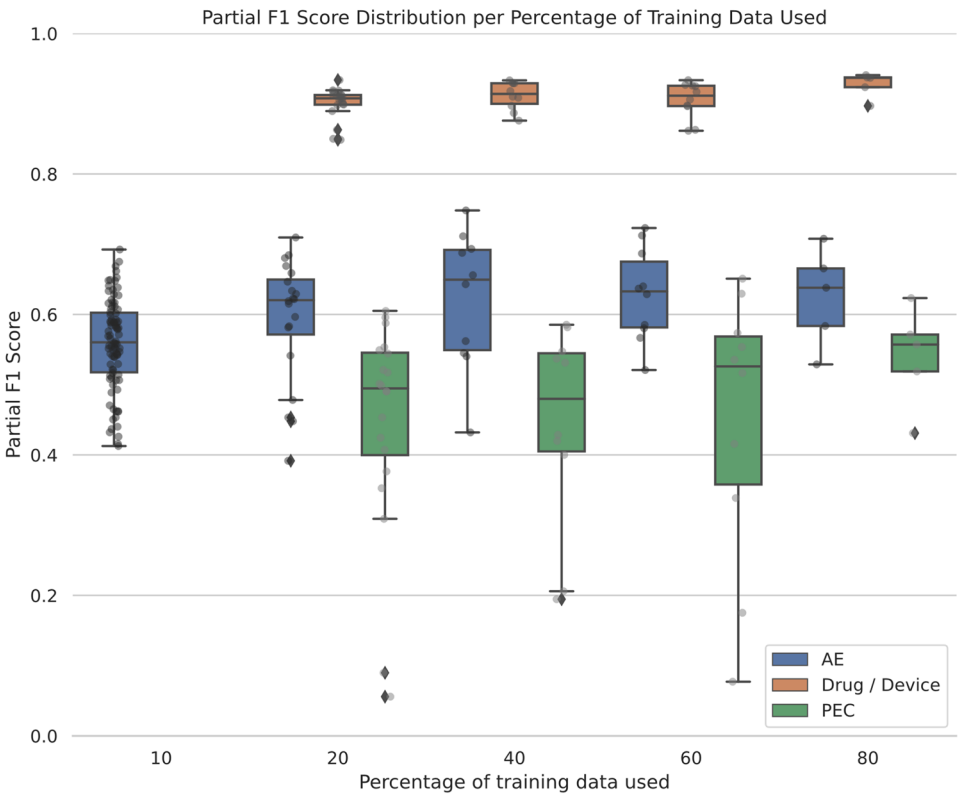

Fig. 6 Boxplots of partial F1 scores of different AE, drug/device, and PEC training data fractions in stratified fivefold cross-validation runs (20% fold, AE: around 320; drug/device: 301; PEC: 284 occurrences). *AE* adverse event, *PEC* Product Event Combination

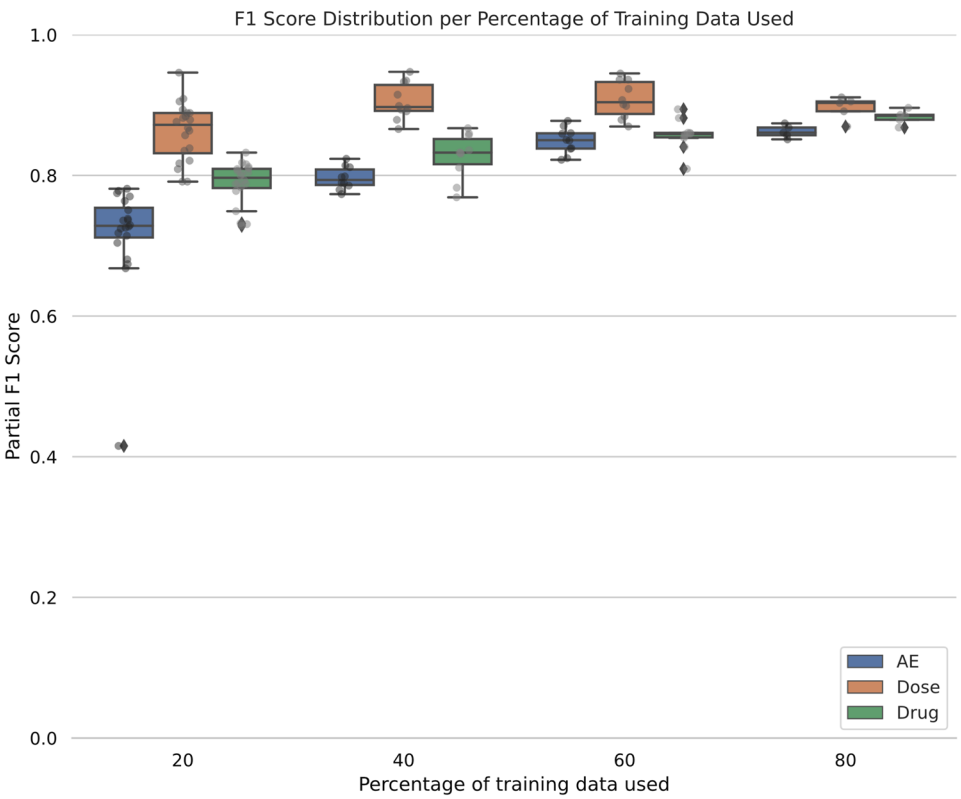

Fig. 7 Boxplots of partial F1 scores of different AE, dose, and drug training data fractions from ADE Corpus in stratified fivefold cross-validation runs (20% fold, AE: around 320; drug/device: 301; dose: 46 occurrences). *ADE* adverse drug effect, *AE* adverse event

Table 6 Results from fractional stratified fivefold cross-validation run per entity from ADE Corpus

Entity	Avg 20%-fold size	Median partial F1 20%	Median partial F1 80%
AE	320	0.728	0.861
Dose	46	0.872	0.903
Drug	301	0.797	0.885

Median partial F1 values for 20% and 80% of training data are listed
ADE adverse drug effect, *AE* adverse event, *Avg* Average

Table 7 Results from fractional stratified fivefold cross-validation run per entity of our dataset retrieved from ADE Corpus

Entity	Avg 20%-fold size	Median partial F1 20%	Median partial F1 80%
AE	141	0.824	0.863
Dose	58	0.907	0.931
Drug/device	118	0.886	0.916

Median partial F1 values for 20% and 80% of training data are listed
ADE adverse drug effect, *AE* adverse event, *Avg* Average

In the next analysis, we investigated fractional stratified k-fold cross-validation for a different dataset with different annotation rules (ADE Corpus V2). To allow comparison of the results with our dataset, the number of occurrences was made similar. We randomly sampled occurrences of AE (n = 1604 of 5742) and drug (n = 1505 of 4927). We used the complete amount of dose data (n = 230 of 230). We executed the 80:20 split on the selected data. The results are shown in Fig. 7 and Table 6. The AE median partial F1 score of the originally annotated ADE Corpus starts and ends at a higher value than the AE values of our dataset (see Tables 5 and 6; 20%: 0.728 vs 0.620; 80%: 0.861 vs 0.638).

We further investigated this effect by executing fractional stratified k-fold cross-validation only with ADE Corpus data from our dataset. In Table 7, the results are shown. Comparing median partial F1 of ADE Corpus AE data in our dataset with the originally annotated ADE data, the start values (20% fold) are higher, although the number of occurrences used are lower (see Tables 6 and 7; 20%: 0.824 vs 0.728; average 20%-fold size: 141 vs 320), while the 80% values are similar (Tables 6 and 7; 80%: 0.863 vs 0.861). The same effect is found for the entity 'drug' (average 20%-fold size: 118 vs 301; median partial F1 20%: 0.886 vs 0.797).

Entity peak performance is reached in our dataset quicker (about 20%) than in the ADE Corpus dataset (about 40%).

The results of the binary classifier experiment are shown in ESM 3, Sect. 2.2 (see the electronic supplementary material).

4 Discussion

We created a new dataset, retrieved from five different sources with 18 biomedical and pharmaceutical entities, which enables ML model entity training independent of other entities. Although human decision-making is considered as the gold standard, in clinical situations, including pharmacovigilance, people make mistakes and do not always agree [4]. Therefore, we established in our annotation process, besides the standard quality measures (e.g., annotation guidelines, IAAs), a 'machine-in-the-loop' approach to increase data completeness and consistency, focusing on false-positive and false-negative model predictions. Although only a small expert team was involved in our annotation process, ML systems found several annotation inconsistencies. In pharmacovigilance, large numbers of case processors and medical experts are involved, which increase the risk of different judgements. In future, the completeness and consistency of Individual Case Safety Reports (ICSRs) may potentially be increased by use of ML systems.

In ESM 3, Sect. 1 (see the electronic supplementary material), we compared our corpus annotation with the ADE Corpus and SMM4H annotation results as an additional and independent quality measure. We found several deviations resulting from differences in the annotation guidelines applied, from different judgements made by annotation teams, or from problems with regard to determining spans consistently with different annotators (e.g., ADE dose annotations: '*high-dose*' in Prominent positive U waves appearing with high-dose intravenous phenylephrine; ADE dose annotation: '*high*' in A case of normotensive scleroderma renal crisis after high-dose methylprednisolone treatment). The intention of our annotation rules was to capture a specific AE as completely as possible. For a medical assessment, it will make a difference if the event appears or disappears (e.g., by changing the dose, withdrawal, or substitution of the medication, challenge/rechallenge, identification of therapeutic response unexpected). These results for SMM4H as well as for the ADE Corpus indicate that different annotation schemes influence the training and test process of an ML system.

In Sect. 3.2, we compared the results of the integrated model with the models trained on single language registers. Our focus was to evaluate the opportunity for production use and to investigate how pronounced the performance of the pure language register models decreases. By using a mixed dataset combining 50% of informal and scientific data, we observed a significantly higher performance decrease with pure informal models compared to pure scientific models. It was not surprising that a model trained on data with a single language register performed better on the test data from the same language register than on test data of the other

language register. But, in addition, we were able to show that an integrated model has higher F1 scores compared to a pure informal model when tested on a pure informal dataset, and slightly higher F1 scores compared to a pure scientific model tested on a pure scientific dataset.

One reason that all models performed overall better in the scientific language register may be due to underlying text sources of the Cleaned English version of Colossal Clean Crawled Corpus (C4.EN corpus) on which the T5-base model was pretrained. In the paper [26], the authors show that elaborated code sources (e.g., patents, Wikipedia, newspapers, scientific literature) were the main source for the model pretraining. Informal language sources seem to be underrepresented. In addition, informal language contains partly offensive terms, 'bad' words, certain demographic identity mentions, or a variety of dialects that probably were excluded by blocklist filtering in the C4.EN corpus [9, 26]. Although formal texts in elaborated codes have a longer, more complicated sentence structure, the use of informal language in our dataset seems to be more diverse, e.g., 'feel and look like a zombie.' Tweets contain information in hashtags and irony and sarcasm (e.g., 'i have been on <drug> for the past few nights and i 've noticed i am slightly drooling a bit . lovely.'). which increases the difficulty for T5 to detect those entities.

ML-based technologies are capable of deriving important insights from the vast amount of data generated every day. The use of ML systems in regulated environments requires CSV. While the underlying CSV requirements largely remain the same, software development activities for ML systems are needed to document evidence that the system is fit for purpose [27, 28]. One important CSV process step is to determine which training data for the validation are sufficient.

In Sect. 3.3, we demonstrated that fractional cross-validation enables investigation of entity performance curve progression. This method can be used to determine entity training data sufficiency in CSV tasks. Testing the performance on entities in our dataset, we identified that for some entities, additional data may be required to reach peak performance (e.g., PTC, comorbidity) or the amount of currently available entity data (medical history/condition) needs to be increased. Our results for fractional stratified fivefold cross-validation showed that the amount of training data needed to successfully fine-tune the model varies for different biomedical entities. The frequency and distribution of the words and concepts in the T5 model pretrained on C4 corpus can affect the stability and variability of the embeddings learned by the model. Identical medical terms (as AE, indication, comorbidity) have different contextual embeddings depending on the surrounding text. Standardized concepts, such as drugs and their intended effects, may be easier for an ML model to detect because they are more consistent in their use across different contexts.

In addition, we investigated the performance of three entities in a subset of data from ADE Corpus with the extraction of this corpus in our dataset. We showed that although the number of occurrences in our dataset is lower than the number of occurrences in the ADE subset, the F1 performance of all entities is better. An easier detection by T5 may be related to different annotation rules, described in Sect. 2.1.3. In contrast to our annotation rules, the ADE Corpus annotation follows a conditional annotation concept: only drug and dose mentions are captured if those entities are related to an AE, which is obviously more difficult to detect by T5.

We demonstrated that this methodology is not specific for our dataset and annotation rules, but can also be used for other dataset and annotation rule combinations (e.g., ADE Corpus).

It is plausible that this methodology could also be used for different transformer models, but this investigation is out of scope for this article. Since the generation of high-quality labeled data is lengthy and expensive, this methodology can be used for evaluation training data sufficiency on the entity level to support the CSV process and may increase regulatory acceptance of ML models and applications in regulated environments.

5 Conclusion

In our paper, we describe the activities for creating a systematically annotated corpus combining, complementing, and harmonizing various corpora with pharmaceutical and biomedical entities based on scientific and informal data. This dataset is made available (excluding internal data) to the research community to train ML models and evaluate the performance of automated methods and systems for entity recognition in unstructured free-text information.

We explored the performance of predictive ML models on different language domain registers. We conclude that comparable performance can be reached by integrated models as compared to single language models when used on the same language type. The integrated model could therefore be considered preferable due to the increased maintenance need when maintaining multiple specialized models for each language type.

We introduced fractional stratified k-fold cross-validation and demonstrated that this methodology enables the investigation of entity performance curve progression and can be used for evaluation training data sufficiency in CSV of ML systems

and may increase regulatory acceptance of ML models and applications.

A future area to advance the dataset could be related to the dataset extension regarding other pharmaceutical entities (e.g., strength), to label additional data (e.g., medical condition/history, comorbidity), to separate entity contents (administration form and packaging), and to incorporate other datasets (e.g., chatbots) to allow extension of language capabilities in future models.

The dataset provided can be used for ML model training or as a part of a shared test dataset for CSV model performance evaluation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-023-01322-3.

Acknowledgements We would like to thank Jayeshkumar Mangroliya, Sasa Lukic, and Peter Jacobi, employees of Syncwork AG for performing parts of Python programming and were responsible for data engineering and data analysis. We would also like to thank Emma Krutzler, Britta Anne Grum for annotation support and John Reinhard Pietzsch, Enrico Santus, Katrin Manlik, Angelo Ziletti, André Hollstein for constructive comments. The MedDRA® trademark is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) on behalf of ICH.

Declarations

Funding Philipp Kazzer, Jayeshkumar Mangroliya, and Sasa Lukic have received compensation for programming and consultancy work from Bayer AG.

Conflict of interest Jürgen Dietrich holds shares in Bayer AG. Jürgen Dietrich is a full-time employee of Bayer AG. Jürgen Dietrich and Philipp Kazzer have no conflicts of interest that are directly relevant to the content of this experiment. The views expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of Bayer AG or Syncwork AG. Jürgen Dietrich has a patent pending on fractional stratified k-fold cross-validation.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The dataset generated during the current experiment is made available in ESM 1, excluding the part retrieved from Bayer's Safety and Product Technical Complaint databases that contains patient private information that cannot be made publicly available.

Code availability The code used for this experiment is not provided, as a patent is pending.

Author contributions JD and PK were involved in the conception and design of the experiments. JD was responsible for dataset annotation, python programming, and model creation, the design and implementation of the data processing, and result interpretation. All authors contributed to the interpretation of the data analysis results and assisted with the concept and draft revisions of the manuscript. All authors reviewed and approved the final manuscript and accept full responsibility for its overall content.

References

1. Kassekert R, Grabowski N, Lorenz D, Schaffer C, Kempf D, Roy P, et al. Industry perspective on artificial intelligence/machine learning in pharmacovigilance. Drug Saf. 2022;45(5):439–48.
2. Working Group XIV Artificial Intelligence in Pharmacovigilance, Minutes of 1st meeting held on 18–19 May 2022 (Geneva). https://cioms.ch/working_groups/working-group-xiv-artificial-intelligence-in-pharmacovigilance/. Accessed 21 Nov 2022.
3. Lewis DJ, McCallum JF. Utilizing advanced technologies to augment pharmacovigilance systems: challenges and opportunities. Ther Innov Regul Sci. 2020;54:888–99.
4. 2nd meeting of the CIOMS Working Group WG XIV on Artificial Intelligence in Pharmacovigilance 10–11 October 2022, Geneva, Switzerland, hybrid meeting. https://cioms.ch/wp-content/uploads/2022/05/CIOMS-WG-XIV-AI-in-PV_2nd-Meeting-minutes_10-11Oct2022.pdf. Accessed 15 Mar 2022.
5. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Minneapolis: Association for Computational Linguistics; 2019. p. 4171–86.
6. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. Minneapolis: Association for Computational Linguistics; 2019. p. 72–8.
7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
8. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task. 2019. p. 21–30.
9. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
10. Petroni F, Piktus A, Fan A, Lewis P, Yazdani M, De Cao N, et al. KILT: a benchmark for knowledge intensive language tasks. Online: Association for Computational Linguistics; 2021. p. 2523–44.
11. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, et al. mT5: a massively multilingual pre-trained text-to-text transformer. Online: Association for Computational Linguistics; 2021. p. 483–98.

12. Raval S, Sedghamiz H, Santus E, Alhanai T, Ghassemi M, Chersoni E. Exploring a unified sequence-to-sequence transformer for medical product safety monitoring in social media. Punta Cana: Association for Computational Linguistics; 2021. p. 3534–46.
13. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform. 2012;45(5):885–92.
14. Dietrich J, Gattepaille LM, Grum BA, Jiri L, Lerch M, Sartori D, et al. Adverse events in twitter-development of a benchmark reference dataset: results from IMI WEB-RADR. Drug Saf. 2020;43(5):467–78.
15. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: a corpus of adverse drug event annotations. J Biomed Inform. 2015;55:73–81.
16. Sarker A, Belousov M, Friedrichs J, Hakala K, Kiritchenko S, Mehryary F, et al. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. J Am Med Inform Assoc. 2018;25(10):1274–83.
17. Weissenbacher D, O'Connor K, Rawal S, Zhang Y, Tsai RT-H, Miller T, et al. Automatic extraction of medication mentions from tweets—overview of the biocreative VII shared task 3 competition. Database. 2023;2023:baac108.
18. Diamantidis N, Karlis D, Giakoumakis EA. Unsupervised stratification of cross-validation for accuracy estimation. Artif Intell. 2000;116(1–2):1–16.
19. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K'in K-fold cross validation. In: 20th European symposium on artificial neural networks, computational intelligence and machine learning (ESANN); 2012: i6doc. com publ; 2012. p. 441–6.
20. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020;27(1):3–12.
21. New Drug Approvals Archive for 2010–2021. https://www.drugs.com/newdrugs-archive/2021.html. Accessed 30 Sept 2021
22. Hugging Face T5 V4.24.0. https://www.huggingface.co/docs/transformers/model_doc/t5#transformers.T5Model. Accessed 16 Nov 2022.
23. Segura-Bedmar I, Martínez Fernández P, Herrero Zazo M. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). 2013: Association for Computational Linguistics; 2013.
24. Opitz J, Burst S. Macro f1 and macro f1. arXiv:1911.03347. 2019. https://doi.org/10.48550/arXiv.1911.03347
25. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013.
26. Dodge J, Sap M, Marasović A, Agnew W, Ilharco G, Groeneveld D, et al. Documenting large webtext corpora: a case study on the colossal clean crawled corpus. In: Proceedings of the 2021 conference on empirical methods in natural language processing. 2021. p. 1286–305.
27. Huysentruyt K, Kjoersvik O, Dobracki P, Savage E, Mishalov E, Cherry M, et al. Validating intelligent automation systems in pharmacovigilance: insights from good manufacturing practices. Drug Saf. 2021;44:261–72.
28. US FDA Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), Discussion Paper and Request for Feedback. 2019. https://www.fda.gov/media/122535/download. Accessed 20 Feb 2023.