Automatic classification of experimental models in biomedical literature to support searching for alternative methods to animal experiments

Mariana Neves (✉ Mariana.Lara.Neves@bfr.bund.de)
Federal Institute for Risk Assessment

Antonina Klippert
Nuvisan (Germany)

Fanny Knöspel
Federal Institute for Risk Assessment

Juliane Rudeck
Federal Institute for Risk Assessment

Ailine Stolz
Federal Institute for Risk Assessment

Zsofia Ban
Federal Institute for Risk Assessment

Markus Becker
Federal Institute for Risk Assessment

Kai Diederich
Federal Institute for Risk Assessment

Barbara Grune
Federal Institute for Risk Assessment

Pia Kahnau
Federal Institute for Risk Assessment

Nils Ohnesorge
Federal Institute for Risk Assessment

Johannes Pucher
Federal Institute for Risk Assessment

Gilbert Schönfelder
Federal Institute for Risk Assessment

Bettina Bert
Federal Institute for Risk Assessment

Daniel Butzke
Federal Institute for Risk Assessment

Research Article

Keywords: alternatives to animal experiments, corpus annotation, text classification, replacement

Posted Date: April 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1526055/v1

RESEARCH

Automatic classification of experimental models in biomedical literature to support searching for alternative methods to animal experiments

Mariana Neves[1*], Antonina Klippert[1,2], Fanny Knöspel[1], Juliane Rudeck[1], Ailine Stolz[1], Zsofia Ban[1], Markus Becker[1], Kai Diederich[1], Barbara Grune[1], Pia Kahnau[1], Nils Ohnesorge[1], Johannes Pucher[1], Gilbert Schönfelder[1,3], Bettina Bert[1] and Daniel Butzke[1]

Correspondence:
Mariana.Lara.Neves@bfr.bund.de
German Centre for the Protection of Laboratory Animals (Bf3R), German Federal Institute for Risk Assessment (BfR), Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Background: European Union legislature requires replacement of animal experiments with alternative methods, whenever such methods are suitable to reach the intended scientific objective. However, searching for alternative methods in the scientific literature is a time-consuming task that requires careful screening of an enormously large number of experimental biomedical publications. The identification of potentially relevant methods, e.g. organ or cell culture models, or computer simulations, can be supported with text mining tools specifically built for this purpose. Such tools are trained (or fine tuned) on relevant data sets labeled by human experts.

Methods: We developed the GoldHamster corpus, composed of 1,600 PubMed (Medline) abstracts, in which we manually identified the used experimental model according to a set of eight labels, namely: "in vivo", "organs", "primary cells", "immortal cell lines", "invertebrates", "humans", "in silico" and "other" (models). We recruited 13 annotators with expertise in the biomedical domain and assigned each article to two individuals. Three additional rounds of annotation aimed at improving the quality of the annotations with disagreements in the first round. Furthermore, we conducted various machine learning experiments based on supervised learning to evaluate the suitability of the corpus for our classification task.

Results: We obtained more than 7,000 abstract-level annotations for the above labels. The inter-annotator agreement (kappa coefficient) varied among labels, and ranged from 0.63 (for "others") to 0.82 (for "invertebrates"), with an overall score of 0.74. The best-performing machine learning experiment used the BioBERT pre-trained model with fine-tuning to our corpus, which gained an overall f-score of 0.82.

Conclusions: We obtained a high agreement for most of the labels, and our evaluation demonstrated, that our corpus is suitable for training reliable predictive models for automatic classification of biomedical literature according to the used experimental models. Our "Smart feature-based interactive" search tool (SMAFIRA) will employ this classifier for supporting the retrieval of alternative methods to animal experiments. The corpus and the source code will be made available.

Keywords: alternatives to animal experiments; corpus annotation; text classification; replacement

1 Introduction

According to Directive 2010/63/EU[1] from the European Union (EU), researchers are only allowed to perform an animal experiment, addressing a particular research question, if no alternative method is already available. Therefore, in the process of obtaining approval for an animal experiment, researchers are required to carry out a comprehensive search to ensure that an alternative method is not yet available. This is a time consuming and complex task that involves many queries to databases with references to scientific publications, and careful screening of candidate publications. There are two important aspects that should be evaluated by researchers when screening for suitable literature: (i) whether the candidate publication's scientific objective is the same as the one that is planned; and (ii) whether the candidate describes an experimental model other than a living (vertebrate) animal (principle of "replacement"). In this work, we focus on the second aspect. To date, there is no tool that supports the search for alternative methods in the literature.

In order to identify whether an experimental approach described in a publication complies with that aspect, the first step is to identify the type of experimental model that is used. While the EU Directive protects all vertebrate animals (mammals, birds, fish, etc.), it does not protect invertebrates, except cephalopods (e.g. octopuses). Therefore, invertebrate animals represent alternative models, and methods based on such animal models are considered alternative methods. Moreover, most in vitro methods (i.e. cell cultures) comply with the replacement principle, and so do experimental computer simulations (in silico methods). So-called ex vivo methods still rely on animals to provide organs or tissues for subsequent experiments, but no live animal experiment is performed, so such methods are not considered animal experiments for the purposes of Directive 2010/63/EU.

The aim of this work is to develop a computational method for automatic classification of publications with respect to the experimental model used. Such a classification is an important part of a tool specialized in finding alternative methods, and it will be included in our corresponding SMAFIRA (SMArt Feature based Interactive RAnking), which is currently being developed in our research group.

Since more than one experimental model can be described in experimental biomedical publications, we tackled the task as a multi-label document classification, i.e.,

[1] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0063`

the automatic assigning of one or more labels to a textual document. Our classification scheme consists of eight labels, and a group of 13 individuals with biomedical expertise annotated a set of 1,600 PubMed[2] abstracts. We chose to rely on PubMed because it is the largest freely available database with references to biomedical publications and it provides Web services for querying and retrieving articles[3]. We experimented with different classifiers to evaluate their ability to classify the experimental models described in the abstracts.

In summary, these are the contributions of this publication:

- The annotation of a novel corpus of scientific abstracts in which we manually assigned the used experimental model. This is also a new benchmark corpus for multi-class, multi-label text classification, a task for which few corpora from the biomedical domain are currently available.
- Python scripts derived from the different machine-learning methods that we experimented with using our corpus, as well as the best performing model (to be made available).

This article has the following structure: Section 2 describes related work for the identification of the used experimental model in publications. In Section 3, we describe the development of our corpus, including the selection of the labels, the retrieval of the documents, and the rounds and quality assessment of annotations. We explain the machine learning methods that we employed in our experiments in Section 4. Section 5 presents the results that we obtained, i.e., statistics of the annotated corpus and the performance of the methods on the corpus. Finally, we discuss some interesting aspects of our corpus and experiments in Section 6.

2 Related Work

Some previous attempts aimed to identify alternative methods (or 3R-relevant methods) in PubMed, i.e. whether a potential alternative to animal experiment was able to replace an animal experiment, reduce the number of animals, or refine the experiments to increase the welfare of animals. Among others, they have relied on curated lists of relevant MeSH (Medical Subject Headings) terms, for instance, to identify publications containing a method based on a cell culture model.

[2] `https://pubmed.ncbi.nlm.nih.gov/`

[3] `http://https://eutils.ncbi.nlm.nih.gov`

ALTBIB[4], i.e. the "bibliography on alternatives to the use of live vertebrates in biomedical research and teaching", uses this approach. However, such predefined search strategies quickly become obsolete as new and potentially relevant MeSH-terms are continuously added to PubMed. In addition to the various MeSH terms that can be used to identify certain classes of experimental methods, there are also two terms specially designed to retrieve alternatives to animal experiments as a whole, namely, "Animal Testing Alternatives" and "Animal Use Alternatives". We have elaborated some case studies of alternatives to animal experiments in the domains of Parkinson's disease, Huntington's disease, breast cancer and stroke before [1] and did not identify the above mentioned specific MeSH-terms in any of the relevant publications. Thus, relevant experimental models and methods may simply be missed when only relying on these two MeSH terms.

A recent work describes the development of a corpus and method for predicting alternatives to animal experiments, or 3R-based literature, based on these two MeSH terms [2]. The authors initially collected around 4,000 citations from PubMed associated with these terms, which was compared to a random set of citations of the same size. They relied on the word2vec algorithm to reveal meaningful patterns, which are then used in a random forest model. However, more details about this research is not yet available, and neither are the data nor the methods.

To the best of our knowledge, there is no manually annotated corpus that can be used for supervised machine learning and the automatic classification of biomedical publications according to used experimental model. However, there are databases that can help with the identification of alternatives to animal experiments. For instance, the Non Animal Technologies (NAT) database[5] provides a collection of non-animal technologies that is available for download from their Web site. The current state of the database comprises 1,068 entries (as of January/2022) and it includes the identification of the type of method or model, e.g. "human studies", "epidemiology", "in silico", "artificial intelligence". Such databases, however, are very limited in their coverage, since collection of data essentially requires human efforts. Moreover, not all entries have a link to a publication, and the database is

[4] `https://ntp.niehs.nih.gov/go/altbib`

[5] `https://www.nat-database.org/`

not transparent about the selection of the proposed methods and who is responsible for this procedure.

Another important collection of non-animal models is the EU reports that the EU Commission regularly releases for some specific topics. As of January/2022, we are aware of four reports for the areas of breast cancer[6], respiratory tract diseases[7], neurodegenerative diseases[8], and immuno-oncology[9]. The corresponding data for each report can be downloaded as a spreadsheet, and similar to the NAT database, it includes information about the methods or models. The collected models, however, are predominantly human-based and thus miss a great portion of candidate alternative models (based upon animal tissues and cells). In addition, these reports describe advanced non-animal methodologies that do not necessarily replace existing animal-based methods.

The annotations in our corpus overlap with some previous corpora that addressed named-entity recognition (NER), for instance, of species [3, 4, 3], anatomical parts [5], or cell lines [6]. However, as opposed to usual NER tasks, we do not aim to identify all mentions of these entities, but focus only on cases that occur in the context of the used model.

Furthermore, not all mentions of vertebrate animals necessarily correspond to an in vivo experiment, i.e. using living animals. If the animal was killed in advance for the removal of organs, tissues, or some cells, such an approach is not considered an animal experiment according to the Directive 2010/63/EU. Within the classification schema we propose, the depicted situation alludes to different labels: "organs" and/or "primary cells". Finally, we are not aware of previously developed NER tools for the extraction of in silico methods, which is one of the labels that we consider.

Comprehensive ontologies and thesauri for the biomedical domain, such as the MeSH terms, address a wide range of concepts related to some of the labels we consider. For example, there are MeSH terms for "primary cell culture" or "cell line". Anyway, these terms do not differentiate between the source of biological materials. The respective labels used in the first version of our corpus designate materials from vertebrate animals only. Other biological materials, used in experimental biomed-

[6] https://op.europa.eu/en/publication-detail/-/publication/8b7b3030-1a64-11eb-b57e-01aa75ed71a1

[7] https://op.europa.eu/en/publication-detail/-/publication/c8ec5086-f988-11ea-b44f-01aa75ed71a1

[8] https://op.europa.eu/en/publication-detail/-/publication/5d9512e7-a89b-11eb-9585-01aa75ed71a1

[9] https://op.europa.eu/en/publication-detail/-/publication/b50a15b5-00ff-11ec-8f47-01aa75ed71a1

ical research, are labeled as "human" or "invertebrate", depending on the source. The basic intention behind our labeling is to allow differentiation between sources and to make a clear distinction between experiments using live animals ("in vivo") and experiments using materials from animals ("ex vivo"). The MeSH-term "animals" does not allow for such a clear distinction. Therefore, we cannot make use of MeSH terms as a straightforward approach to identify the used experimental model. Indeed, a preliminary analysis of the correlation of our labels with the respective MeSH terms concluded that a simple mapping between them is not feasible (cf. Section 6.5).

In one previous attempt, our group developed an ontology for this field [7], to support the Go3R search engine [8]. Go3R was one of the first attempts to develop a tool for finding alternative methods to animal experiments. Its ontology was divided in 28 branches and contained more than 16,000 concepts. The tool utilized a maximum entropy algorithm, which was trained on a collection of 3,000 manually annotated documents, to predict whether an article was 3R-relevant. However, the tool is no longer available, and the methods, the training data, and the ontology were never released.

3 GoldHamster Corpus

We annotated 1,600 from PubMed abstracts to support a classification according to the experimental models they use. In this section, we describe: (a) how we defined the annotation scheme, (b) the queries we used to search for abstracts, and (c) details of the annotation process.

3.1 Definition of the annotation schema

Labels for annotation were designed to help basically distinguishing experimental research with live vertebrate animals ("in vivo", i.e. animal experiments by legal definition) from research using materials from vertebrate animals having been sacrificed in advance ("organs/tissues", "primary cells", i.e. no animal experiments by legal definition), or from research being conducted with no need of using any vertebrate laboratory animal at all ("immortal cell lines", "invertebrates", "human", "in silico"). The labels with a short description are listed in Table 1.

In particular, the label "in vivo" refers to experiments using living vertebrates (and cephalopods), i.e. equal to animal experiments. The labels "organs" and "pri-

mary cells" refer to experiments with biological materials from vertebrate animals that were killed in advance (i.e. no animal experiments by legal definition). The label "immortal cell lines" refers to experiments using immortalized cell lines of vertebrate animals that can be ordered from cell and tissue collections, e.g. the American type culture collection (ATCC). The label "invertebrates" refers to experiments with invertebrates (e.g. flies and "worms") or invertebrate material (excluding cephalopods, i.e. no animal experiment by legal definition). The label "human" refers to experiments with humans or human material. "In silico" was included to indicate research using computer simulations. We labeled any observational (but not experimental) study as "other", e.g. clinical retrospective studies.

Table 1 List of labels in the GoldHamster corpus. The annotators could assign one or more label to an abstract, or even no label at all.

Label	Description
in vivo	Experiments in living vertebrates (and cephalopods)
organs	Vertebrate organs and tissues
primary cells	Vertebrate primary and stem cells
immortal cell lines	Vertebrate immortalized and cancer cell lines
human	Experiments with humans or human material
invertebrates	Experiments with invertebrates or invertebrate material (excluding cephalopods)
in silico	Computer simulations
others	Other experiments, e.g., retrospective studies and meta-analyses

3.2 Retrieval of abstracts

Complex search strategies using MeSH (Medical Subject Headings) were devised to retrieve the abstracts and searches were performed in PubMed/MEDLINE (on August 20, 2019). The queries consisted of combinations of MeSH-terms referring to certain experimental models and techniques (e.g. "Animals, Genetically Modified[MeSH] OR Animal Experimentation[MeSH] OR ...") and relevant categories (e.g. "Diseases Category[MeSH]") with filters (e.g. "English[lang]"). Such combinations then were extended to target certain clusters of MEDLINE-abstracts. The terms which were searched for cluster headings were "in vivo", "organs and tissues", "primary and stem cells", "immortalized and tumor cells", "in silico", "invertebrates", "humans", "other" (cf. Section 1 in the supplementary material). From every retrieved list, the first 200 abstracts were downloaded and were included in the corpus. We created eight queries, i.e., one for each cluster, and they yielded a

combined corpus of 1,600 abstracts. We provide our queries in the supplementary material (Section 1).

3.3 Annotation process

A group of 13 annotators carried out the annotation using the TeamTat tool[10] [9]. All annotators have a doctoral degree in the field of biomedical science or are currently research assistant in our department. The annotation guidelines is provided in the supplementary material (Section 2). The annotation process was carried out in two rounds: (i) round 1, namely "r1", in which each of the 1,600 abstracts was annotated by two annotators, who were randomly selected; (ii) round 2, which was split into three small rounds, namely "r2.1", "r2.2", and "r2.3", in which a selected team of the annotators resolved some of the disagreements in annotations from the first round. We describe both rounds in details below.

First round - r1. We arranged the 1,600 abstracts in 40 units of 40 abstracts each and assigned them to annotators. Each annotator received a set of four (i.e., 160 abstracts) to eight (i.e., 320 abstracts) units. The annotators were required to highlight the applied experimental models, which are described in either the title or text of the abstract. Even though we addressed the problem as a text classification task, highlighting a text span was necessary because TeamTat does not support document-level annotation. However, for the sake of simplicity, we asked the annotators to highlight only one mention (text span) for each label, instead of all mentions of the experimental model in the text. In addition, we did not specify which text should be highlighted, since only the labels were relevant. The annotators were encouraged to consult external resources in the Web, such as Cellosaurus[11] [10] or ATCC[12], for the identification of the origin of a cell line. Finally, if the abstract did not allude to any experimental model, the annotators were asked not to assign a label to it.

Second rounds - r2.1, r2.2, and r2.3. In the second rounds (cf. 5.1), five selected annotators (based on "reliability", i.e. high average agreement with the other annotators) resolved some of the disagreements in annotations from the first round.

[10] `https://www.teamtat.org/`

[11] `https://web.expasy.org/cellosaurus/`

[12] `https://www.atcc.org/`

This was a pilot study that considered 333 abstracts[13] to assess whether we could improve the inter-annotator agreement with additional rounds of annotation. The selected annotators had to consider which of the two (anonymous) annotators in the first round was correct. We assigned the second round annotators to abstracts that he or she did not annotate in the first round. If the annotator conceived a third opinion about the labels, the abstract was flagged to be removed, since no agreement between any two annotators was obtained. This phase consisted of three short rounds in which we selected some particular disagreement combinations of labels from the first round:

- r2.1: 71 abstracts in which one annotator assigned only the label "invertebrate", while the other one assigned something else, i.e. one label other than "invertebrate" or multiple labels (which could also include "invertebrate");
- r2.2: 114 abstracts in which one annotator assigned only the label "human", while the other one assigned something else, i.e. one label other than "human" or multiple labels (which could also include "human");
- r2.3: 148 abstracts in which one annotator assigned only the label "in vivo", while the other one assigned something else, i.e. one label other than "in vivo" or multiple labels (which could also include "in vivo").

4 Classifiers

In this section, we describe the classifiers we trained to automatically assign labels to abstracts. We experimented with the current state-of-the-art approach for the biomedical domain, i.e. fine-tuning a pre-trained BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) model [11] with our annotated corpus, as well as Support Vector Machines (SVM). Since BioBERT-based approaches are time consuming, we relied on SVM for the preliminary experiments that aimed at assessing the various datasets, annotations (with and without agreement), and two possible sets of labels.

Our main classifier was based on BioBERT[14], which is a pre-trained language representation model that achieved state-of-the-art results for many biomedical natural language processing (NLP) tasks (e.g. text classification, question answering,

[13] of the 701 articles without a full agreement, cf. more details in Section 5.1

[14] `https://github.com/dmis-lab/biobert`

and named entity recognition) without requiring substantial modification in the model architecture. BioBERT is based on the BERT language model [12], but word representations were learned from PubMed abstracts and PubMed Central (PMC) full texts, in addition to the Wikipedia articles and books that were originally used to train BERT. We relied on the implementation of BERT in the Transformers model[15] for TensorFlow. In our experiments, we used the Adam optimizer, epsilon of 1×10^{-8} and decay of 0.01. We experimented with various values to adjust the model hyperparameters, such as learning rate, batch size, and the number of epochs, as later described in Section 5.2.

Furthermore, we also tried the SVM classifier as implemented in the Python scikitlearn library[16]. We tried the four kernels functions[17] available in the library, namely linear, polynomial, RBF, and sigmoid. All experiments utilized a TF-IDF representation of the abstracts and we ran various experiments in order to achieve the best parameters for each of the kernel functions (cf. Section 5.2).

5 Results

In this section, we present an analysis of the manual annotations that we obtained for the GoldHamster corpus, and the results of the experiments with automatic prediction of the labels.

5.1 Corpus analysis

After the first round of annotation, in which two annotators screened each of the articles, we obtained 7,737 annotations[18]: 1,970 for "in vivo", 1,397 for "human", 1,171 for "invertebrates", 892 for "others", 740 for "organs", 663 for "in silico", 455 for "primary cell lines", and 449 for "immortal cell lines". These values are the total of annotation with duplicates, i.e. including situations in which two annotators agreed in assigning a certain label to an abstract. In addition, in 190 cases, an annotator did not assign any of the labels to an abstract. From the 1,600 documents, 899 had a full agreement, i.e. exactly the same sets of labels were assigned by both

[15] https://github.com/huggingface/transformers

[16] https://scikit-learn.org/stable/modules/svm.html

[17] https://scikit-learn.org/stable/modules/svm.html#svm-kernels

[18] The upper limit for any label thus is 2 x 1,600 (abstracts) = 3200, if a label was assigned to all abstracts in the corpus by both allotted annotators.

annotators. From the remaining documents without full agreement, we reviewed 333 in the additional rounds (r2.1, r2.2, and r2.3).

Table 2 Statistics of the corpus in terms of the number of abstracts per label. We present statistics for all labels (cf. 3.1), three rounds of annotation (cf. 3.3) and when considering all annotations (All, left side) or only the one for which two annotators agree (Agree, right side). We also consider the "cell lines" label, which is a fusion of the "primary cell lines" and "immortal cell lines" labels. The comparison for the rounds in terms of equality (=), increase (△), or decrease (▽) is with respect to the previous column (round), i.e. after the addition of the corresponding round. The values do not include duplicates, i.e. a label is only counted once if there is an agreement for it.

Labels	Round 1 All/Agree	+ Round 2.1 All/Agree	+ Round 2.2 All/Agree	+ Round 2.3 All/Agree
invertebrates	263/193	14▽/16△	=/=	2▽/=
in_vivo	483/346	1▽/2△	=/1△	19▽/21△
in_silico	216/128	4▽/4△	2▽/3△	1▽/8△
human	325/169	1△/=	37▽/36△	4▽/2△
organs	261/130	3▽/=	=/=	47▽/13△
immortal_cell_lines	160/56	2▽/=	23▽/5△	2▽/6△
primary_cells	189/60	6▽/=	1▽/=	22▽/6△
others	402/131	23▽/5△	25▽/35△	18▽/8△
cell_lines	312/114	8▽/=	24▽/5△	24▽/12△
none	165/25	11▽/=	=/=	=/=
total docs	1,600/1,189	=/23△	=/67△	=/36△

We summarize the number of labels that we obtained after each of the annotations rounds in Table 2[19]. We show the impact of each additional round as compared to the previous column. Furthermore, we present results when considering all annotations (hereafter called "All") and when only considering the ones for which there was an agreement between two annotators (hereafter called "Agree"). As expected, when full agreement is considered, the number of articles with any annotation reduces, i.e., from 1,600 to 1,189 for the first round. However, the latter increases after the annotation of the additional rounds, namely to: 1,279 (r1 + r2.1 + r2.2), 1,248 (r1 + r2.1 + r2.3), 1,292 (r1 + r2.2 + r2.3), and 1,315 (r1 + r2.1 + r2.2 + r2.3). For all rounds, the number of documents without any annotation remained equal to 25.

In the three additional rounds, the selected annotators received the anonymous annotations from the two annotators in the first round and were required to decide, which annotation is the correct one. If no first round annotation was judged correct,

[19] After removing duplicates for cases in which two annotators agreed.

respective abstracts were removed from the corpus. We removed the following number of abstracts from each additional round: 11 (from 71) in r2.1, 19 (from 114) in r2.2, and 23 (from 148) in r2.3. We present the list of PMIDs in the supplementary material (section 3).

Table 3 Agreement in terms of Cohen's (κ) between annotators for the corpus in each round (cf. 3.3), for the all labels (cf. 3.1), and for the overall corpus. An agreement is moderate if the (κ) is higher than 0.6, and strong if higher than 0.8.

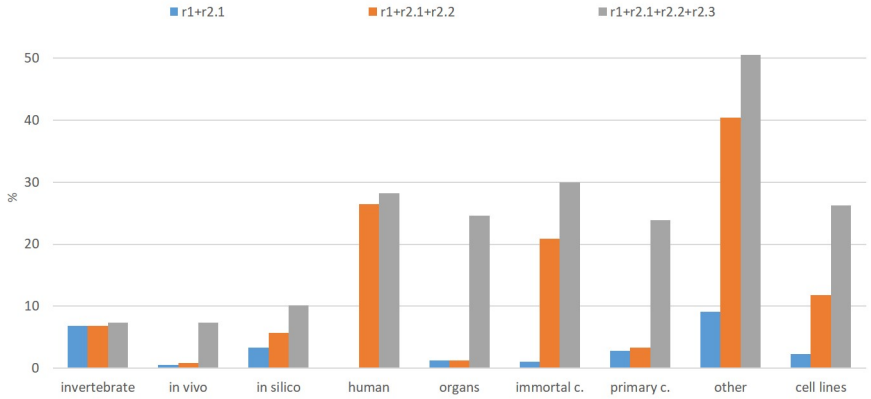
Labels	r1	r1+r2.1	r1+r2.2	r1+r2.3	r1+r2.1+r2.2+r2.3
invertebrates	0.82	0.88	0.82	0.83	0.88
in vivo	0.78	0.78	0.78	0.83	0.84
in silico	0.72	0.74	0.73	0.75	0.79
human	0.63	0.63	0.80	0.64	0.81
organs	0.62	0.63	0.62	0.77	0.78
immortal cell lines	0.49	0.50	0.59	0.54	0.64
primary cell lines	0.45	0.46	0.45	0.54	0.56
others	0.42	0.46	0.55	0.46	0.63
cell lines	0.59	0.60	0.64	0.66	0.74
overall	0.62	0.64	0.67	0.67	0.74

In Table 3 we show the level of agreement between the annotators with respect to individual labels in terms of the kappa coefficient (κ) [13]. Regarding the first round, we obtained a strong agreement for "invertebrates" (0.82), moderate agreement for "in vivo" (0.78), "in silico" (0.72), "human" (0.63), and "organs" (0.62), and weak agreement for "immortal cell lines" (0.49), "primary cells" (0.45), and "others" (0.42).

Two of the lowest agreements were for the "primary cell lines" and "immortal cell lines" labels, which are indeed difficult to distinguish (if not using the Cellosaurus). Therefore, we explored merging the two labels into one: the "cell lines" label. As expected, the kappa coefficient for the merged label is higher (0.59) than the ones from the respective single labels (0.49 and 0.45, respectively).

Figure 1 shows the increase of the Cohen's kappa coefficient, when compared to relying only on round r1, for each label, for all annotation rounds, and the many combination of these. As expected, the improvement was higher when considering all annotations of the second round, since it includes a higher number of articles (and their annotations). Regarding each of the additional rounds, a considerable improvement for the "invertebrates" label occurred in r2.1, as for the "human" label in r2.2, and for the "in vivo" label in r.2.3. These labels were the focus of the

Figure 1 Increase (in percentage) on the kappa coefficient for each annotation round. The improvement is with respect to the original annotations (round 1), for each original label as well as the combined label "cell_lines".



respective rounds. However, other labels also had a significant improvement, such as "organs", "primary cells", "immortal cell lines", and "others".

5.2 Evaluation

We ran various experiments to evaluate the suitability of our corpus for predicting the labels. All results are in terms of the standard metrics of precision, recall, and f-score.

The additional rounds of annotation aimed to improve the corpus with the addition of each new round. For the sake of comprehensiveness, we used all possible combinations of the first round (r1) with the subsequent rounds (r2.1, r2.2, r2.3) for machine learning, and evaluated the influence of each combination on the f-scores. Here are the eight combinations of rounds that we considered: (i) r1; (ii) r1+r2.1; (iii) r1+r2.2; (iv) r1+r2.3; (v) r1+r2.1+r2.2; (vi) r1+r2.1+r2.3; (vii) r1+r2.2+r2.3; (viii) r1+r2.1+r2.2+r2.3.

In addition, for each of the above combinations of annotation rounds, either we considered all annotations (All), or only the ones with agreement between annotators (Agree). Furthermore, we tested two sets of labels, containing either the two original cell line-based labels (i.e. "immortal cell lines" and "primary cells"), or only the merged one: "cell line" (CL). Here are these four situations:

- (i) All: all annotations, independent of the agreement;

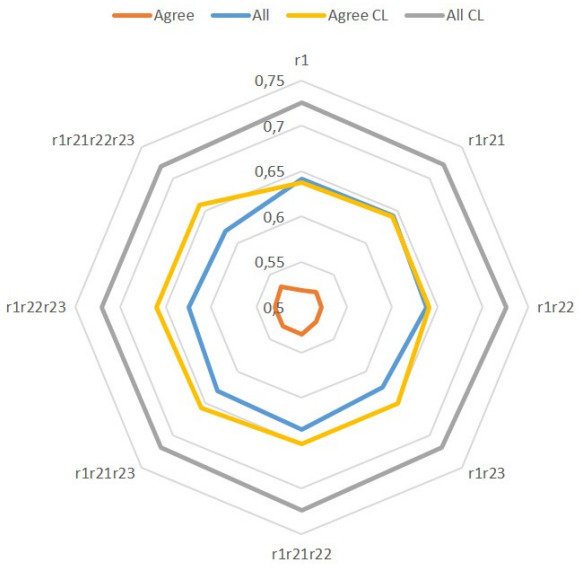
- (ii) All+CL: like (i) , but with the addition of the merged "cell lines" label, instead of the two separate labels "primary cells" and "immortal cell lines";
- (iii) Agree: only the annotations with agreement of two annotators;
- (iv) Agree+CL: like (iii), but with the addition of the merged "cell lines" label, instead of the two separate labels "primary cells" and "immortal cell lines".

This resulted in 32 combinations: 8 round sets x 2 agreement levels x 2 label sets. Finally, in order to perform a 10-fold cross validation, we split the collection of abstracts into 10 parts in a stratified way, i.e., in a way to obtain datasets with a similar distribution of labels as in the complete corpus (cf. statistics in Section 4 of the supplementary material).

Our first experiments aimed at examining which of the above combinations obtained better results. We relied on a linear SVM model and chose a TF-IDF representation of abstracts. Figure 2 depicts the f-scores that we obtained for all above mentioned combinations. All experiments that relied on the merged cell line label obtained higher results (averages of 0.65 and 0.73, for "Agree CL" and "All CL", respectively) than the ones using the two separate cell line labels (averages of 0.53 and 0.64, for "Agree" and "All", respectively). In most cases, however, there seem to be no significant difference between the various combinations of rounds. The only exception was seen when considering the combination of all annotations and the two separate cell line labels (i.e. the "All" curve in Figure 2). In this particular case, the results decreased with the additional rounds. Finally, for both set of labels, considering all annotations (All), and not only the ones with agreement (Agree), obtained a considerable better performance, i.e.: averages of 0.64 over 0.53, when relying on two separate cell line labels; and 0.73 over 0.65, when relying on the merged cell line label (CL). Since improvement for the merged cell line label was substantial, we did not consider the two separate cell labels in any of our subsequent experiments.

We compared the various SVM kernel functions implemented in Python scikitlearn library. We started by considering the default parameters of the kernel functions for the "Agree" and the "All" annotations. For the four kernel functions, i.e. linear, polynomial, radial basis function (RBF), and sigmoid, the results were always superior for the "All" annotations. Therefore, we considered the "All" annotation in our subsequent experiments with SVM and ran 10-fold cross validation experiments for

Figure 2 Variation of the f-score for each round combination, agreement level, and set of labels. Higher results occurred when considering all annotations (not only the ones with an agreement) and the merged label for cell lines (CL). Substantial difference between the various datasets was only observed for the "All" curve. All experiments refer to the SVM linear algorithm.



all eight datasets, and for a set of values for the hyperparameters which are required by each kernel function. We tried the following values for each kernel function:

- RBF: "gamma" of 0.25, 0.5, 0.6, 0.75, 0.8, 1 and 2
- sigmoid: "coef0" of -1, -0.5, -0.25, 0, 0.25, 0.5 and 1
- polynomial: "degree" of 2, 3 and 4; "coef0" of -100, -50, -10, -1, 0, 1, 10, 50, 100

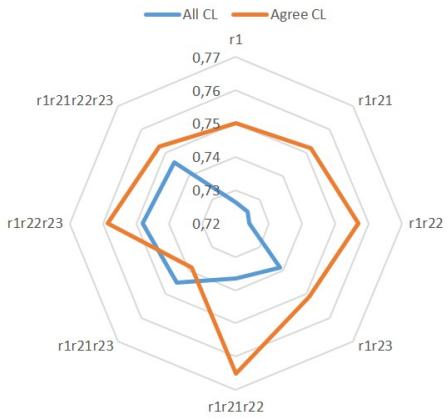
The best performances obtained by each kernel function are summarized in Table 4 and the best hyperparameters were the following: "gamma" of 0.75 for RBF, "coef0" of -0.25 for sigmoid, and "degree" of 2 and "coef0" of 1 for polynomial. For all SVM algorithms, the best results were obtained when considering all annotations (All). For each kernel, the best results occurred for different datasets, namely: r1 for linear and RBF, r1+r2.3 for sigmoid, and r1+r2.2 for polynomial.

For the BioBERT approach, we ran 10-fold cross validation experiments for all combinations (cf. Figure 3) and for both the "All" and "Agree" annotations. For these experiments, we used the following initial hyperparameters: learning rate of 5×10^{-5}, batch size of 32, and 10 epochs. The best results occurred for the rounds

Table 4 Performance of the methods for the prediction of the labels. We show the best f-scores that we obtained for each method over all datasets and for a 10-fold cross-validation. We highlight in bold the best results for each label and in general.

Labels	SVM				BioBERT
	linear	RBF	sigmoid	polynomial	
invertebrates	0.9032	0.7183	0.8686	0.8714	**1.0**
in_vivo	0.8027	0.6885	0.7833	0.8047	**0.8235**
human	0.7410	0.6228	0.7163	0.7044	**0.8696**
organs	0.6371	0.2297	0.5578	0.6248	**0.6667**
cell_lines	0.5770	0.4161	0.5704	0.5704	**0.6667**
in_silico	0.7935	0.3094	0.6937	0.7504	**0.8**
others	0.6264	0.5726	0.6365	0.6302	**0.9333**
All (average)	0.7258	0.5082	0.6895	0.7080	**0.8228**

Figure 3 Variation of the f-score for the various combinations and for the "All CL" and "Agree CL" annotations. The highest f-score occurred for the "Agree" annotations and for r1+r2.1+r2.2 dataset. The values of f-score are the average for the 10-fold cross validation. All experiments refer to the BioBERT method.



r1+r2.1+r2.2 with the "Agree" annotations, as opposed to the results that we initially obtained for the SVM classifiers.

We ran further experiments to assess the set of hyperparameters that obtained the highest performance. For these runs, we considered only the "Agree" annotations from the rounds r1+r2.1+r2.2, and the best performing split (number 3) of the 10-fold cross validation. The sets of values for the hyperparameters that we considered were the following: learning rate of 1×10^{-5}, 5×10^{-5}, 1×10^{-4}, and 5×10^{-4}; batch size of 16 and 32; and epochs of 10, 20, 30, 40, and 50. We show the results for the 40 experiments that we ran with all combinations of hyperparameters in the supplementary material (Section 5). The best set of hyperparameters that we

obtained was the following: learning rate of 1×10^{-4}, batch size of 16, and 20 epochs. Table 4 shows the overall f-score, as well as the f-scores for all labels, for the best performing BioBERT model, which outperformed all SVM classifiers for all labels.

6 Discussion

Here we discuss some issues related to our annotations and experiments, such as possible reasons for the inconsistencies in the annotation process, the contributions of the different annotation rounds to the results, and additional experiments with additional features such as MeSH terms and sections of abstracts.

6.1 Annotation process

After the first round of annotations, in which 13 annotators independently reviewed 1,600 abstracts with two-fold overlap (to determine inter-annotator agreement), we performed three additional rounds with subsets of first round annotations to partly resolve disagreements from the first round. While we obtained an improvement in the inter-annotator agreement in all additional rounds (cf. Table 3), we did not always observe an improvement in the predictions after training SVM or BioBERT-based models with the respective sets of annotations (cf. Figure 2).

As expected, agreement between annotators decreased with the assignment of more than one label per article. For articles where the annotator assigned only one label, we found a full agreement of 71%. This value decreased to 36% when considering two labels per article, and to 12.5% when considering three labels per article.

We identified the labels (and their combinations) for which an inter-annotator agreement was obtained. Table 5 presents the number of articles with such agreement after the first round of annotation. As expected, full agreement occurred mostly in abstracts assigned to just one label. Furthermore, full agreement of annotations was found in 31 abstracts labeled twice, and only six times when three labels were present.

For abstracts without full agreement, we obtained 167 combinations of disagreements. The most frequent disagreement that we observed was the assignment of the label "human" by one annotator, and the label "others" by the other one (n = 55). Indeed, many of the abstracts assigned to "others" were retrospective studies in which patients were involved. Thus, according to our guideline, the label "others"

Table 5 Number of the 899 articles with full agreement according to the obtained labels. The table is divided into three groups: documents with one, two, or three labels. "Total" is the total number of documents in each group, while "No. Docs" is the number of documents with each of the labels (or set of labels).

	Total	No. Docs	Labels
1 label	862	213	"in vivo"
		156	"invertebrates"
		136	"human"
		97	"others"
		92	"organs"
		85	"in silico"
		30	"immortal cell lines"
		28	"primary cell lines"
		25	none
2 labels	31	6	"in silico+in vivo"
		5	"in vivo+primary cell lines"
		4	"in silico+invertebrates", or "immortal cell lines+in vivo", or "in vivo+organs"
		2	"organs+primary cell lines", or "in silico+primary cell lines", or "human+invertebrates", or "in silico+others"
3 labels	6	1	"human+in vivo+primary cell lines", or "in vivo+invertebrates", or "immortal cell lines+primary cell lines", or "in vivo+others", or "in silico+organs", or "immortal cell lines+organs"

was the correct one. We provide more details about the disagreements in Section 6 of the supplementary material.

6.2 Correlation with the initial database queries and the corresponding MeSH terms

Our initial assumption was that a query based on MeSH terms could not precisely distinguish between certain types of experimental models[20], since such terms are too indistinct for this specific task, i.e. to distinguish "animal experiments" from "experiments using animals" (cf. Section 2).

The list of 1,600 PMIDs in our corpus results from eight database queries, which were designed to roughly represent the labels in our schema. From each of the hit lists, we retrieved the top 200 PMIDs for which an abstract was available. After the annotation of the abstracts, we assessed whether the assigned labels in our corpus correspond to our initial queries.

[20]in particular "in vivo" models and models using biopsies or tissues from animals killed in advance ("ex vivo")

Figure 4 Correlation of the labels (in the graph) with the queries (in the legend) used for retrieving the abstracts. The values represent the percentage of articles assigned to the particular label that correlated with the query. The labels were derived from the "Agree" annotations and the rounds r1+r2.1+r2.2+r2.3.

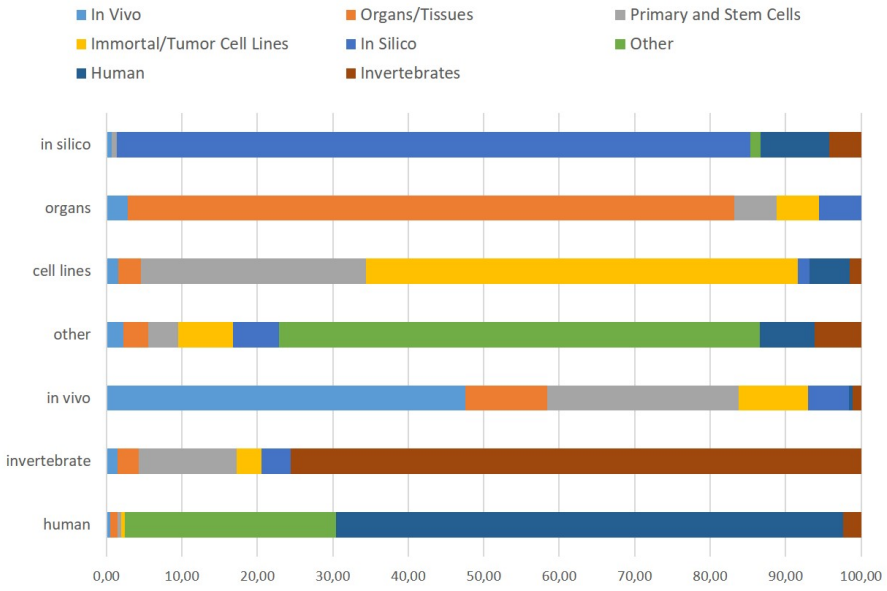
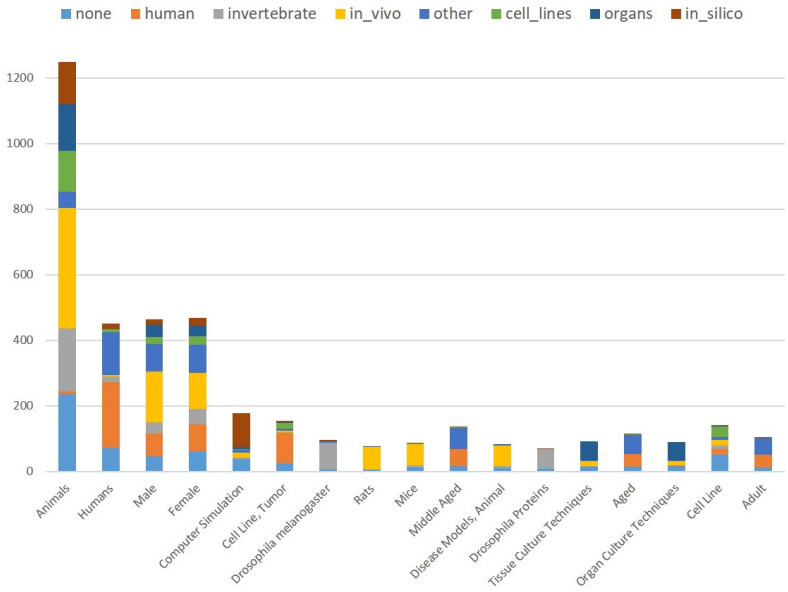


Figure 4 presents the percentage of abstracts that agree with the respective label that was assigned in the rounds r1+r2.1+r2.2+r2.3 (i.e. full agreement only). The "organs" and "in silico" labels obtained the highest agreement (over 80%). Furthermore, these values are even higher than the highest f-scores obtained for these labels in our machine learning experiments: 0.6667 and 0.8, respectively. The agreement for "invertebrate" was high (75.6%) but lower than the f-score for this label, which reached up to 1.0 (cf. Table 4). The lowest agreement was the one for "in vivo": only 47.57%.

Further, we assessed the potential of relying on MeSH terms to support the prediction of the labels. We retrieved all MeSH terms associated with our 1,600 abstracts and obtained 2,828 distinct terms when considering only the ones associated to at least ten documents. Only 22 of the abstracts contained no MeSH terms at all. Then we analyzed the correlation between the terms and the labels by checking the MeSH terms that occur in abstracts with a particular label. We summarize in Figure 5 the correlation for the MeSH terms[21] to a certain label. On the one hand, some

[21] with a frequency of at least 50

Figure 5 Association of the labels and some MeSH terms. The values in the vertical axis represent the number of times that the term was associated to an abstract of the respective label. Only MeSH terms which occur more than 50 times were considered.



terms had a clear correlation with only one of our labels and are good discriminators for our labels, e.g., "Computer Simulation" for "in silico", "Cell Line, Tumor" for "human", "Drosophila melanogaster" for "invertebrates", and "Rats" and "Mice" for "in vivo". On the other hand, some terms frequently occur for more than one label, and are probably not good discriminators, such as "Animals", "Humans", "Male", and "Female". Therefore, building reliable queries based on MeSH terms is a challenge for most of the labels that we address.

6.3 Contribution of the additional rounds of annotation

The additional rounds (r2.1, r2.2, and r2.3) aimed at achieving a higher agreement for the annotations. Indeed, the inter-annotation agreement (kappa score) shown in Table 3 confirms an increase in the agreement for all labels. However, this procedure also resulted in the removal of some abstracts for which no agreement was possible.

Our preliminary experiments with the SVM classifiers for all 32 combinations of annotation rounds, sets of labels, and annotation agreement (cf. Figure 2) showed that a higher overall f-score is obtained when considering all annotations, and not only the ones with agreement. However, subsequent experiments with BioBERT (cf.

Figure 3) showed that relying on annotations with a higher agreement could indeed increase the performance.

While the additional rounds brought an improvement in the kappa scores for all labels (cf. Figure 1), we cannot state the same for the performance in the prediction of these labels (cf. Section 7 in supplementary material). The round 2.1 aimed at improving the agreement for the "invertebrates" label. While a higher f-score for this label was achieved with SVM linear, it did not occur for BioBERT, which actually suffered a decrease in the f-score. Round 2.2 focused on the "human" label. We could observe a higher f-score for this label for SVM linear, but just a slightly higher one for BioBERT. Finally, round 2.3 cleared disagreements for the "in vivo" label. We observed a decrease in the f-score for this label for SVM linear, and a slightly increase in the f-score for BioBERT.

In summary, as depicted in the figures in Section 5 of the supplementary material, it is not possible to observe a clear correlation between the additional rounds of annotation and an increase in the performance of the labels. However, as depicted in Table 3, the additional rounds improved the quality of our corpus, i.e. increased the kappa coefficient. Therefore, they are important to assure the quality of the annotations.

6.4 Prediction of multiple labels

We evaluated the ability of the model to predict multiple labels for an abstract. For a training data of 1183 abstracts, 82 of them contain multiple labels, when considering "Agree" annotations only. In comparison, this number increases to 440 abstracts with multiple labels when considering "All" annotations. Because the "All" datasets contain more annotations, models trained on these annotations are able to predict more abstracts with multiple labels. For a test dataset containing 65 abstracts, we obtained 23 abstracts with multiple labels, when trained with "All" annotations, as opposed to only 7 when training with the "Agree" annotations.

We also observed a variation in the models' ability to predict multiple labels depending on the training dataset. The average number of abstracts with multiple labels vary significantly: from 0.8 (r1 dataset) to 2.7 (r1+r22+r23 dataset) for the "Agree" annotations, and from 9.9 (r1+r22+r23 dataset) to 15.4 (r1+r21 dataset) for the "All" annotations. However, we did not observe any clear trend, i.e. that a

higher inter-annotator agreement always generated more (or less) multiple labels. However, training models only on "Agree" annotations will certainly cause that fewer abstracts will be assigned to multiple labels. Furthermore, when relying on the BioBERT model, we observed that the f-scores for models trained on "Agree" annotations only are usually higher than the ones obtained with "All" annotations (cf. Figure 3). However, this is because, with BioBERT, the precision is much higher for models trained with "Agree" annotations, while the recall is usually lower than the ones obtained for the "All" annotations. Moreover, a lower recall directly affects the ability of a model to predict multiple labels. However, when aiming at a higher recall, models trained on "All" annotations should be preferred.

6.5 Additional semantic features

We assessed whether additional semantic features, e.g. MeSH terms, could boost the performance in automatic prediction of classes (experimental models), since these terms reflect the important content of published articles. However, some of these terms might refer to information only present in the full text of articles. We did not consider full text in our annotation process and experiments, but only abstracts.

We ran experiments with all MeSH terms originally assigned to the articles, as well as MeSH term subsets based on pre-defined thresholds, i.e., considering only the ones that occur in at least a certain number of articles. There were 5,158 distinct MeSH terms in our corpus of 1,600 articles. When considering the values of n = 10, 25, 50, and 100 for the threshold, i.e. occurrence of MeSH terms in "n" articles, the number of distinct terms reduced to 298, 93, 38, and 15, respectively. In the supplementary material (Section 8), we depict the variation in the number of terms per article for each of these thresholds.

We added the MeSH terms, in addition to the title and abstract text, in our machine learning experiments by concatenating the terms at the beginning of the text. We ran experiments for all the above threshold values, as well as for the best performing models that we obtained for SVM linear and BioBERT (cf. Table 4). Table 6 presents the results. For the SVM linear classifier, the f-score increased with the addition of the MeSH terms. No substantive difference occurred for the thresholds 0, 10, 25 and 50, but the performance decreased considerably for the threshold 100. For the BioBERT classifier, just a slight improvement arose for the

threshold 100, otherwise, the results were always lower than the original one without the terms.

Table 6 Prediction of the labels with or without (w/o) considering the MeSH terms. We considered various thresholds for the minimum frequency of the terms (in relation to the number of articles in which they appear).

	w/o MeSH	thresholds				
		0	10	25	50	100
SVM linear	0.6928	0.7235	0.7346	0.7299	0.7330	0.7096
BioBERT	0.8228	0.7449	0.7617	0.8086	0.7968	0.8325

In a second experiment, we evaluated whether the abstract structure, i.e. the identification of the sections (background, methods, results, and discussion), had a beneficial impact on the performance of the classifiers. For these experiments we considered either the originals sections in PubMed or the ones predicted by the ArguminSci tool [14], which is the tool that best performed for this task in our previous evaluation [15]. We provide more details in Section 9 in the supplementary material. We considered many sections and combinations of sections. However, none of these experiments outperformed the best results for neither SVM linear nor BioBERT previously shown in Table 4.

6.6 Limitations and Future Work

For this first version of our corpus, we designed a set of eight labels. While these labels cover most important vertebrate models in experimental biomedical research, the resolution of human models is underdeveloped. Therefore, we plan to expand the set of labels with classes referring to human models, i.e. “human in vivo”, “human organs/tissues”, “human cell lines”. Furthermore, we plan to improve our guidelines and examples regarding the distinction of (vertebrate or human) “primary cells” and “immortal cell lines”. Annotators will receive adequate training and will be obliged to use the Cellosaurus [10] for the identification of the cell lines. There also will be more emphasis on the hierarchical position of label “others” which actually is at the top of a hierarchical system, and captures abstracts that describe biomedical research that uses other than experimental approaches, e.g. observational studies.

We carried out annotation only on abstract-level, and even though the annotators highlighted some text passages in the TeamTat annotation tool, we did not require them to highlight all mentions of a particular model. As future work, we plan

to consider these annotations for further experiments, e.g., for a sentence-based classification.

We did not check all abstracts with disagreements. From 1,600 abstracts, of which 701 did not achieve full agreement between annotators, we reviewed only 333 of them in the additional rounds, leaving 368 for future revision. While it is clear that the agreement (kappa score) increased with the additional rounds of annotations, our experiments have not always shown that these additional rounds resulted in an increase in the performance (f-score) for predicting the labels. In spite of this, we also plan to add new rounds of annotations for these remaining articles to achieve a corpus of abstracts that is more representative/relevant for the real world set of abstracts, i.e. the MEDLINE database. The evaluation set based on our "Agree" annotations may be biased towards abstracts that require simple classification skills only, leaving out complicated ones.

We integrated the best performing BioBERT model in our SMAFIRA Web tool, which is currently under development. SMAFIRA is a search engine that aims at supporting researchers when searching for alternative methods to animal experiments. Our BioBERT classifier provides a real-time classification of the abstracts which are automatically retrieved from PubMed based on the user input query. This allows the user to filter the list of results based on the one or more of our labels.

7 Conclusion

In this publication, we presented a new corpus of 1,600 Medline abstracts and manually annotated it using a set of eight labels (in vivo, organs, primary cells, immortal cell lines, human, invertebrates, in silico, and others). We obtained more than 7,000 annotations (multi-labeling was possible) and involved 13 annotators in the process (two annotations per document). The annotations that we obtained achieved at least a moderate agreement for almost all labels (after all additional rounds), and even a high agreement for some of them. We expect that this corpus can support the development of applications in the field of alternative methods to animal experiments, as well as serve as a benchmark for biomedical text classification tasks.

We ran many machine learning experiments to assess the feasibility of using our corpus to predict the labels. These experiments aimed at the identification of the best set of labels, the best annotations (with our without full agreement), the im-

pact of each additional annotation round, and the best set of parameters for the various algorithms. The best result was obtained by BioBERT when relying on annotations with full agreement. Therefore, we plan to perform an additional round of annotation for the remaining abstracts which still contain disagreements.

We provided an adequate analysis of our annotations and of the disagreements between the annotators. These insights will certainly be valuable when designing the next version of our corpus. Further, we investigated the correlation of our annotations with respect to the original queries (cf. Section 6.2), which were used to retrieve the abstracts, and the MeSH terms associated to the articles. These results support our claim that queries based on MeSH terms are not adequate for precisely identifying the experimental method (or model) described in a publication.

Acknowledgements

None.

Funding

This work was funded by the Federal Institute for Risk Assessment, Berlin....

Abbreviations

None.

Availability of data and materials

The dataset with the annotations was uploaded as supplementary material....

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

D.B. designed the label scheme and compiled the corpus. M.N. prepared the corpus, analyzed the data, and carried out the experiments. A.K., F.K., J.R., A.S., Z.B., M.B., K.D., B.G., P.K., N.O., J.P., and D.B. annotated the corpus. G.S. and B.B. coordinated the work. M.N. and D.B. wrote the main manuscript text. All authors reviewed the manuscript.

Authors' information

Not applicable.

Author details

[1]German Centre for the Protection of Laboratory Animals (Bf3R), German Federal Institute for Risk Assessment (BfR), Berlin, Germany. [2]Current affiliation: Nuvisan ICB GmbH, Müllerstraße 178 13353 Berlin, Germany. [3]Institute of Clinical Pharmacology and Toxicology, Charité - Universitätsmedizin Berlin, Charitéplatz 1 10117 Berlin, Germany.

References

1. Butzke, D., Dulisch, N., Dunst, S., Steinfath, M., Neves, M., Mathiak, B., Grune, B.: SMAFIRA-c: A benchmark text corpus for evaluation of approaches to relevance ranking and knowledge discovery in the biomedical domain. Research Square (2020). doi:10.21203/rs.3.rs-16454/v1. `https://doi.org/10.21203/rs.3.rs-16454/v1`
2. Ritskes-Hoitinga, M., Alkema, W.: The use of artificial intelligence for the fast and effective identification of three rs-based literature. Alternatives to Laboratory Animals **0**(0), 02611929211048447 (0). doi:10.1177/02611929211048447. PMID: 34581190. https://doi.org/10.1177/02611929211048447
3. Pafilis, E., Frankild, S.P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Arvanitidis, C., Jensen, L.J.: The species and organisms resources for fast and accurate identification of taxonomic names in text. PLOS ONE **8**(6), 1–6 (2013). doi:10.1371/journal.pone.0065390
4. Gerner, M., Nenadic, G., Bergman, C.M.: Linnaeus: A species name identification system for biomedical literature. BMC Bioinformatics **11**(1), 85 (2010). doi:10.1186/1471-2105-11-85
5. Pyysalo, S., Ananiadou, S.: Anatomical entity mention recognition at literature scale. Bioinformatics **30**(6), 868–875 (2013). doi:10.1093/bioinformatics/btt580. https://academic.oup.com/bioinformatics/article-pdf/30/6/868/17344635/btt580.pdf
6. Kaewphan, S., Van Landeghem, S., Ohta, T., Van de Peer, Y., Ginter, F., Pyysalo, S.: Cell line name recognition in support of the identification of synthetic lethality in cancer from text. Bioinformatics **32**(2), 276–282 (2015). doi:10.1093/bioinformatics/btv570. https://academic.oup.com/bioinformatics/article-pdf/32/2/276/6690131/btv570.pdf
7. Wächter, T., Sauer, U., Doms, A., Grune, B., Alvers, M., Spielmann, H., Schroeder, M.: An ontology to represent knowledge on animal testing alternatives. Nature Precedings (2009). doi:10.1038/npre.2009.3148.1
8. Sauer, U.G., Wachter, T., Grune, B., Doms, A., Alvers, M.R., Spielmann, H., Schroeder, M.: Go3r - semantic internet search engine for alternative methods to animal testing. ALTEX - Alternatives to animal experimentation **26**(1), 17–31 (2005). 2020-09-09T15:04:58.000Z - JCR autoupdate
9. Islamaj, R., Kwon, D., Kim, S., Lu, Z.: TeamTat: a collaborative text annotation tool. Nucleic Acids Research **48**(W1), 5–11 (2020). doi:10.1093/nar/gkaa333. https://academic.oup.com/nar/article-pdf/48/W1/W5/33433452/gkaa333.pdf
10. Bairoch, A.: The cellosaurus, a cell-line knowledge resource. Journal of biomolecular techniques : JBT **29**(2), 25–38 (2018). doi:10.7171/jbt.18-2902-002. 29805321[pmid]
11. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2019). doi:10.1093/bioinformatics/btz682. https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf
12. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). doi:10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
13. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012). 23092060[pmid]
14. Lauscher, A., Glavaš, G., Eckert, K.: ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In: Proceedings of the 5th Workshop on Argument Mining, pp. 22–28. Association for Computational Linguistics, Brussels, Belgium (2018). doi:10.18653/v1/W18-5203. https://aclanthology.org/W18-5203
15. Neves, M., Butzke, D., Grune, B.: Evaluation of scientific elements for text similarity in biomedical publications. In: Proceedings of the 6th Workshop on Argument Mining, pp. 124–135. Association for Computational Linguistics, Florence, Italy (2019). doi:10.18653/v1/W19-4515. https://www.aclweb.org/anthology/W19-4515

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GoldHamsterSuppMaterial.pdf