Journal of Biomedical Informatics: X 4 (2019) 100058

Contents lists available at ScienceDirect

Journal of Biomedical Informatics: X

journal homepage: www.journals.elsevier.com/journal-of-biomedical-informatics-x

Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations

Anna Koroleva[a,b,*], Sanjay Kamath[a,c], Patrick Paroubek[a]

[a] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
[b] Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
[c] LRI Univ. Paris-Sud, CNRS, Université Paris-Saclay, F-91405 Orsay, France

ARTICLE INFO

Keywords:
Trial outcomes
Semantic similarity
Natural Language Processing
Deep learning
Pre-trained language representations
Spin detection

ABSTRACT

Background: Outcomes are variables monitored during a clinical trial to assess the impact of an intervention on humans' health.Automatic assessment of semantic similarity of trial outcomes is required for a number of tasks, such as detection of outcome switching (unjustified changes of pre-defined outcomes of a trial) and implementation of Core Outcome Sets (minimal sets of outcomes that should be reported in a particular medical domain).

Objective: We aimed at building an algorithm for assessing semantic similarity of pairs of primary and reported outcomes.We focused on approaches that do not require manually curated domain-specific resources such as ontologies and thesauri.

Methods: We tested several approaches, including single measures of similarity (based on strings, stems and lemmas, paths and distances in an ontology, and vector representations of phrases), classifiers using a combination of single measures as features, and a deep learning approach that consists in fine-tuning pre-trained deep language representations.We tested language models provided by BERT (trained on general-domain texts), BioBERT and SciBERT (trained on biomedical and scientific texts, respectively).We explored the possibility of improving the results by taking into account the variants for referring to an outcome (e.g.the use of a measurement tool name instead on the outcome name; the use of abbreviations).We release an open corpus with annotation for similarity of pairs of outcomes.

Results: Classifiers using a combination of single measures as features outperformed the single measures, while deep learning algorithms using BioBERT and SciBERT models outperformed the classifiers.BioBERT reached the best F-measure of 89.75%.The addition of variants of outcomes did not improve the results for the best-performing single measures nor for the classifiers, but it improved the performance of deep learning algorithms: BioBERT achieved an F-measure of93.38%.

Conclusions: Deep learning approaches using pre-trained language representations outperformed other approaches for similarity assessment of trial outcomes, without relying on any manually curated domain-specific resources (ontologies and other lexical resources). Addition of variants of outcomes further improved the performance of deep learning algorithms.

1. Introduction

Outcomes in clinical research are the variables monitored during clinical trials to assess how they are affected by the treatment taken or by other parameters. Outcomes are one of the most important elements of trial design: they represent the objectives of the trial; the primary outcome (the main monitored variable) is used to determine the trial's statistical power and to calculate the needed sample size.

There are several data sources that contain information on trial outcomes. First, outcomes of clinical trials are recorded in trial registries - open online databases that store information on planned, ongoing or completed research. Second, outcomes are defined in protocols of clinical trials. Last, outcomes are presented in texts of medical research articles, where they can occur in two main types of contexts: 1) definition of outcomes that were assessed in the trial ("*Primary outcome will be **overall survival**.*") – context similar to that in protocols; and 2) reporting of results for an outcome ("*Patients of the treatment condition showed significantly greater reduction of **co-morbid depression** and*

* Corresponding author.
E-mail addresses: koroleva@limsi.fr, aakorolyova@gmail.com (A. Koroleva).

https://doi.org/10.1016/j.yjbinx.2019.100058

Available online 17 October 2019

anxiety *as compared to the waiting list condition.*"). We will refer to the outcomes occurring in the first type of contexts as *pre-defined outcomes*, and to the outcomes occurring in the second type of context as *reported outcomes*.

A number of tasks require comparing two outcomes (from the same or different sources) to establish if they refer to the same concept.

First of all, assessing similarity between pairs of outcomes is vital to detect outcome switching. Outcomes should normally be clearly defined before the start of a trial, usually at the moment of the first registration [1,2], and should not be changed without a justification. Consistency in trial outcome definition and reporting is essential to ensure reliability and replicability of a trial's findings and to avoid false positives based on reporting only the variables that showed statistically significant results confirming the researchers' hypothesis. Despite the widely acknowledged importance of proper reporting of outcomes, outcome switching – omitting pre-defined outcomes of a trial or adding new ones – remains a common problem in reporting clinical trial results. The COMPare Trials project [3,4] showed that, in 67 assessed trials, 354 pre-defined outcomes were not reported, while 357 outcomes that had not been defined in advance were added to the trial's report. Outcome switching can occur at several points: pre-defined outcomes in a medical article may be changed compared to those recorded in trial registry/protocol; reported outcomes in an article may differ compared to those recorded in trial registry/protocol or to those pre-defined in the article.

Outcome switching is directly related to two well-known problems of medical research reporting: bias, i.e. choosing only the outcomes supporting the trial hypothesis [5–7], and spin, i.e. reporting only favourable outcomes and thus making research results seem more positive than the evidence justifies [8–13]. Spin in clinical trials assessing an intervention poses a serious threat to the quality of health care: clinicians reading trial reports with spin tend to overestimate the effects of the intervention studied [14]. Besides, spin in research articles causes spin in health news coverage and press releases [15,16], that can raise unjustified positive expectations regarding the intervention among the public.

Checking an article for outcome switching is a part of assessment for bias and spin. The checks can be performed at several levels: the outcomes recorded in the corresponding trial protocol/registry entry should be compared to the primary and secondary outcomes defined in the article; the pre-defined primary and secondary outcomes (in the protocol/registry and in the article) should be compared to the outcomes reported in the article. To perform all these comparisons, it is necessary to assess pairs of outcomes for their semantic similarity.

Another task that requires comparing outcomes concerns the core outcome sets (COS) - agreed minimum sets of outcomes to be measured in trials in particular domains[1]. The core outcome set for a domain that a trial belongs to should be compared to the outcomes defined in a trial protocol/registry entry, to identify gaps in the trial planning at an early stage and improve the trial design. Besides, the COS can be compared to the article reporting a trial to check if results for all the core outcomes are reported.

In this paper, we propose an approach to measuring semantic similarity between phrases referring to outcomes of clinical trials. It is important to note that an outcome is a complex notion that is characterized by several aspects:

- outcome name: "***depression severity***";
- measurement tool used if the outcome cannot be measured directly: "*depression severity measured by* ***the Beck Depression Inventory-II (BDI-II)***",
- time points at which the outcome is measured: "*differences in the Symptom Index of Dyspepsia* ***before randomization***, ***2 weeks*** *and* ***4 weeks after randomization***, *and* ***1 month*** *and* ***3 months after completing treatment***";
- patient-level analysis metric, e.g., change from baseline, final value, time to event: "***change from baseline*** in body mass index (BMI)" population-level aggregation method, e.g. mean, median, proportion: "*the* ***mean*** *number of detected polyps*", "***the proportion of patients*** suffering from postoperative major morbidity and mortality";
- type of analysis of results based on the population included, i.e. intention-to-treat analysis (all the enrolled patients are analyzed, even those who dropped out) or per-protocol analysis (only the patients who followed the protocol are analyzed): "*the change in IOP from baseline to week 4 at 8 a.m. and 4 p.m.* ***for the per protocol (PP) population*** *using a "worse eye" analysis*";
- covariates that the analysis of the outcome is adjusted for: "*whole body bone mineral content of the neonate, adjusted for* ***gestational age*** *and* ***age at neonatal DXA scan***";
- reasons for using a particular outcome (explanation of relevance, references to previous works using the outcome): "*the physical and mental component scores (PCS and MCS) of the Short Form 36 (SF-36),* ***a widely used general health status measure***".

Outcome mentions necessarily contain the outcome name or the measurement tool name, which are used to refer to the outcome. However, all the other items are not mandatory. The level of detail in an outcome mention can differ between different data sources: e.g. registry outcomes tend to be longer and described in more detail than those defined in the articles. Thus, an inherent problem for establishing the similarity between two outcomes is comparing detailed outcome descriptions to under-specified ones. Besides, it is questionable whether two outcomes differing in e.g. type of analysis (intention-to-treat vs per-protocol) are different outcomes or different aspects of the same outcome. In this work, we consider two outcomes to refer to the same concept if the outcome/measurement tool names of the two are the same, disregarding the other aspects.

To the best of our knowledge, automatic outcome similarity assessment has not been addressed yet. We present the first corpus of sentences from biomedical articles from PubMed Central (PMC)[2] annotated for outcomes and their semantic similarity. This corpus has been created in the context of a project aimed at automating spin detection in clinical articles, which is a part of the Methods in Research on Research (MiRoR) programme[3], an international multi-disciplinary research project aiming at reducing the waste in biomedical research.

We propose deep learning methods using pre-trained language representations to evaluate similarity between pairs of outcomes. We compare a number of representations, pre-trained on general-domain and on domain-specific datasets. We compare the deep learning approach to some simple baseline similarity measures.

2. Related work

The previous work distinguished between the notions of semantic similarity and semantic relatedness. Pedersen and colleagues [17] define relatedness as "the human judgments of the degree to which a given pair of concepts is related", and state that it is a more general concept of semantics of two concepts, while similarity is a type of relatedness, usually defined via the "is-a" relation between the concepts in a taxonomy or ontology. Measuring semantic similarity of clinical trial outcomes has not been addressed as a separate task before, but semantic similarity and relatedness assessment and paraphrase recognition attracts substantial attention as it is required in a wide range of domains and applications. Similarity is measured between long or short texts or concepts. Measures used are often based on specialized

[1] http://www.comet-initiative.org/glossary/cos/.

[2] https://www.ncbi.nlm.nih.gov/pmc/.

[3] http://miror-ejd.eu/.

lexical resources (thesauri, taxonomies). In this section, we provide on overview of several works on similarity and relatedness in the biomedical domain.

The measures of semantic similarity and relatedness can be divided into the following groups: string similarity measures, path-based measures, information content-based measures, and vector-based measures. Similarity and relatedness can be measured on different levels: word, term, concept, or sentence.

2.1. String similarity measures

String-based similarity measures are the simplest similarity measures based only on the surface form of the compared phrases, without taking into account the semantics. Still, they find their use in measuring the semantic similarity in the biomedical domain, e.g. the work of Soganciogl u and colleagues [18] used, among other measures of similarity, a number of string-based measures: q-gram similarity (the number of q-grams from the first string over the q-grams obtained from the other string), block distance (the sum of the differences of corresponding components of two compared items), Jaccard similarity (the number of common terms in two sets over the number of unique terms in them), overlap coefficient (the number of common terms in two sets divided by the size of the smaller set), and Levenshtein distance (the minimum number of changes required to transform one string into another).

2.2. Ontology-based measures

2.2.1. Path-based measures

Ontologies contain a formal, structured representation of knowledge. A number of similarity measures based on paths between the concepts in ontologies exist, such as the path similarity (the shortest path connecting the concepts in the hypernym–hyponym taxonomy); the Leacock-Chodorow similarity score [19] (the shortest path connecting the concepts and the maximum depth of the taxonomy used); the Wu-Palmer similarity score [20] (the depth of the senses of the concepts in the taxonomy and that of their most specific ancestor node); a metric of distance in a semantic net, introduced by Rada and colleagues [21], calculated as the average minimum path length between all combinations of pairs of nodes corresponding to concepts; the minimum number of parent links between the concepts [22]. The most commonly used ontology in the general domain is WordNet [23], however, similarity measures based on general-domain resources are stated to be ineffective for domain-specific tasks [17]. A number of works proposed to adapt the existing measures of semantic similarity, which are based on WordNet, to the biomedical domain using the available medical ontologies, in particular SNOMED CT[4], MeSH[5] (Medical Subject Headings), or the Gene Ontology [21,17,24,25,22,18]. Importantly, when similarity is assessed on the sentence level, tools such as Metamap [26] are needed to map the sentence text to concepts from the Unified Medical Language System (UMLS) [18]. Metamap finds both words and phrases corresponding to medical concepts, which makes this approach more reliable that assuming that each word is a concept.

2.2.2. Information content-based measures

Information content (IC) reflects the amount of information carried by a term in a discourse. The notion of IC was introduced by Resnik [27] who proposed to measure the IC of a concept as $IC(c) = -logp(c)$, where c denotes a concept and p(c) denotes the probability of the concept c occurring in a corpus. IC can be used to measure the similarity of two concepts by calculating the amount of information shared by them. Resnik [27] proposed to measure the similarity of concepts as the IC of their least common subsumer (the most specific taxonomical ancestor of the two terms).

IC-based similarity measures have been used in the biomedical domain. Pedersen and colleagues [17] assessed IC-based measures introduced by Resnik [27] and Lin [28] on a set of pairs of medical terms. Sánchez and Batet [29] proposed an overview of IC-based similarity measures (e.g. [27,28]) and developed a method of computing IC from the taxonomical knowledge in biomedical ontologies, in order to propose new IC-based semantic similarity measures. Benaouicha and Hadj Taieb [30] proposed to measure semantic similarity based on IC, using topological parameters of the MeSH taxonomy.

A notable work of Harispe and colleagues [31] provides a more systematic view at ontology-based similarity measures. The authors analyzed a number of ontology-based semantic similarity measures to assess whether some of the existing measures are equivalent and which measures should be chosen for a particular application. The authors classify the similarity measures into a few categories: edge-based measures (similarity of two concepts is calculated according to the strength of their interlinking in an ontology); node-based measures, divided into feature-based approaches (evaluating a concept by a set of features made of its ancestors) and approaches based on information theory (similarity of concepts is calculated according to the amount of information they provide, as a function of their usage in a corpus); and hybrid approaches, combining edge-based and node-based approaches.

Apart from representing the compared concepts, ontologies can be used to exploit contextual features to assess the similarity of new terms. Spasic and Ananiadou [32] proposed to represent the context of a term by syntactic elements annotated with information retrieved from a medical ontology. The sequences of contextual elements are compared using the edit distance (number of changes needed to transform one sequence into another).

2.3. Vector-based measures

Distributional models of semantics, representing term information as high-dimensional vectors, are successfully used in a number of tasks, including semantic similarity assessment (e.g. [33]). In the biomedical domain, Soganciogl u and colleagues [18] used distributed vector representations of sentences built with the word2vec [34] model to compute sentence-level semantic similarity. Henry and colleagues [35] compared a number of multi-word term aggregation methods of distributional context vectors for measuring semantic similarity and relatedness. The methods assessed include summation or mean of component word vectors, construction of compound vectors using the compoundify tool (a part of the Perl word2vec interface package[6]), and construction of concept vectors using MetaMap. None of the evaluated multi-word term aggregation methods was significantly better than the others. Park and colleagues [36] developed a concept-embedding model of a semantic relatedness measure, combining the UMLS and Wikipedia as an external resource to obtain contexts texts for words not presented in the UMLS. Concept vector representations were built upon the context texts of the concepts. The degree of relatedness of concepts was calculated by the cosine similarity between corresponding vectors. This approach is stated to overcome the issue of limited word coverage, which the authors state to pose problems for earlier approaches.

2.4. Methods combining several measures

Some approaches combine several of the above-listed measures of similarity and/or relatedness. Soganciogl u and colleagues [18] developed a supervised regression-based model combining the string similarity measures, ontology-based measures, and distributed vector

[4] http://www.snomed.org/.

[5] https://www.nlm.nih.gov/mesh/meshhome.html.

[6] https://sourceforge.net/projects/word2vec-interface/.

representations as features. Henry and colleagues [37] developed an approach combining statistical information on co-occurrences of UMLS concepts with structured knowledge from a taxonomy, based on concept expansion using hierarchical information from the UMLS.

The common feature of the majority of the listed approaches to semantic similarity assessment is the use of domain-specific resources such as ontologies, that require laborious curation. Recently, Blagec and colleagues [38] suggested an alternative approach to evaluating semantic similarity of sentences from biomedical literature. The authors employed neural embedding models that are trained in an unsupervised manner on large text corpora without any manual curation effort needed. The models used in this work were trained on 1.7 million PubMed articles. The models were evaluated on the BIOSSES dataset of 100 sentence pairs [18]. The unsupervised model based on the Paragraph Vector Distributed Memory algorithm showed the best results, outperforming the state-of-the-art results for the BIOSSES dataset. The authors also proposed a supervised model including string-based similarity metrics and a neural embedding model. It was shown to outperform the existing ontology-dependent supervised state-of-the-art approaches.

3. Existing datasets

A few datasets annotated for semantic similarity of biomedical concepts or texts exist. Pedersen et al. [17] were the first to introduce a set of 30 pairs of medical terms annotated for semantic relatedness by 12 annotators on a 10-point scale.

Pakhomov and colleagues [39] created a set of 101 medical term pairs that were rated for semantic relatedness on a 10-point scale by 13 medical coding experts. The set was initially compiled by a practicing Mayo Clinic physician.

Pakhomov and colleagues [40] compiled a set of 724 pairs of medical terms from the UMLS, belonging to the categories of disorders, symptoms and drugs. The dataset included only concepts with at least one single-word term, to control for impact of term complexity on the judgements on similarity and relatedness. Further, a practicing physician selected pairs of terms for four categories: completely unrelated, somewhat unrelated, somewhat related, and closely related. Each category comprised approximately 30 term pairs. The pairs were rated for semantic similarity and relatedness by 8 medical residents.

The BIOSSES dataset [18] contains 100 pairs of sentences selected from the Text Analysis Conference Biomedical Summarization Track Training Dataset. The sentence pairs were rated for similarity on a 5-point scale by five human experts.

Wang and colleagues [41] aimed at creating a resource for semantic textual similarity assessment in the clinical domain. The authors assembled MedSTS, a set of 174,629 sentence pairs from a clinical corpus at Mayo Clinic. Two medical experts annotated a subset of 1,068 sentence pairs with similarity scores in the range from 0 to 5.

Table 1 summarizes the characteristics of the existing datasets.

4. Annotation of outcome pairs

For us the application of interest is detection of spin related to incorrect reporting of the primary outcome in abstracts of articles reporting randomized controlled trials (RCTs), in particular, omission of the primary outcome. This task is very specific and requires a corpus with annotations for semantic similarity of pairs of primary and reported outcomes. The task of semantic similarity assessment of outcomes can be regarded as a subtask of semantic similarity assessment of medical term pairs, which has been explored in previous works and for which a few datasets exist. However, there is an inherent difference between a corpus of outcome pairs and the existing corpora of medical term pairs: while the existing corpora of medical term pairs contain terms belonging to different categories (e.g. drugs, symptoms and disorders), all the terms in a corpus of outcome pairs belong to the same

Table 1
Existing datasets annotated for semantic similarity/relatedness in the biomedical domain.

Paper	Similarity/relatedness	Items	Number of pairs	Scale	Selection process	Number of annotators	Competence of annotators
Pedersen et al. [17]	Relatedness	Medical terms	30	1–10	Manual selection	12	Physicians and medical coders
Pakhomov et al. [39]	Relatedness	Medical terms	101	1–10	Manual selection	13	Medical coding experts
Pakhomov et al. [40]	Similarity and relatedness	Medical terms	724	0–1600 (pixel offsets)	Two-step (automated + manual)	8	Medical residents
Soganciogl u et al. [18]	Similarity	Sentences	100	0–4	Manual selection	5	Unspecified
Wang et al. [41]	similarity	sentences	1068	0–5	Automatic selection	2	Medical experts

class (outcomes, i.e. measures or variables). In a corpus containing several categories, it can be expected that the items of the same category are judged to be more similar to each other than to the items of other categories (e.g. all the drug names are more similar to each other than to the names of disorders), while in a corpus with a single category this criterion does not apply. The relation of semantic similarity is simpler for outcomes: two outcome mentions are either same (refer to the same measure/variable), or different, hence the relation is binary and can be annotated on a 0–1 scale. On the contrary, in the existing corpora multi-item scales were necessary to annotate similarity/relatedness (drugs names are more similar to each other than disorder names, but the level of similarity within the category of drug names vary).

As no corpus with annotation for semantic similarity of outcomes exists, we created and annotated our own, that we release as a freely available dataset [42]. It is based on a set of 3,938 articles from PMC[7] with the publication type "Randomized controlled trial". The corpus annotation proceeded in two steps: annotation of primary and reported outcomes, and annotation of semantic similarity between them. As it proved to be impossible to recruit within a reasonable time frame several annotators with sufficient level of expertise in the domain of medical research reporting, the annotation work was performed by one single annotator with expertise in NLP, trained and consulted by three experts in clinical research reporting.

4.1. Annotation of outcomes

The annotation and extraction of primary and reported outcomes is the subject of a separate paper, here we only present in brief the annotation principles that are important for the topic of this paper.

For primary outcome annotation, we aimed at annotating contexts that explicitly define the primary outcome of a trial, e.g.:

*"We selected **the shortened version of the Chedoke Arm & Hand Activity Inventory (CAHAI-7)** as the primary outcome measure."*

To find these contexts, we randomly selected 2,000 sentences that contain the word *"primary"* or its synonyms, followed by the word *"outcome"* or its synonyms, with the distance no more than 3 token between them. The synonyms of the words "*primary*" and "*outcome*" used in sentence selection are shown in Table 2. The sentences were selected from full-text articles. We annotated the longest continuous text span that includes all the relevant information about the trial's outcome, such as measurement tool used, time points, etc.

For reported outcomes annotation, we selected the Results and Conclusions sections of the abstracts of the articles for which we previously annotated the primary outcomes. 1,940 sentences constituted the corpus for reported outcomes annotation.

Reporting outcomes are characterized by high diversity: they can be expressed by a noun phrase, a verb phrase or an adjective. The same outcomes can be reported in different ways, e.g. the following sentences report the same outcome:

1. *"At 12-month follow-up, the intervention group showed a significant positive change (OR = 0.48) in **receiving information on healthy computer use** compared to the usual care group."*
2. *"The intervention group showed a significant positive change (OR = 0.48) in **receiving information on healthy computer use** at 12-month follow-up, compared to the usual care group."*
3. ***"Receiving information on healthy computer use** in the intervention group showed a significant positive change (OR = 0.48) at 12-month follow-up, compared to the usual care group."*

In different variants of the sentence, it is possible to annotate as the outcome either:

[7] https://www.ncbi.nlm.nih.gov/pmc/.

Table 2
Synonyms of the words "*primary*" and "*outcome*" used in sentence selection.

Word	Synonyms
Primary	Main, first, principal, final, key
Outcome	Endpoint/end-point/end point, measure, variable, assessment, parameter, criterion

1. *"change (OR = 0.48) in receiving information on healthy computer use"*,
2. *"receiving information on healthy computer use at 12-month follow-up"*, or
3. *"Receiving information on healthy computer use"*.

However, it appears reasonable to have the same outcome annotated in all of the variants. Thus, we annotated the shortest possible text span for reported outcomes.

4.2. Annotation of semantic similarity of pairs of outcomes

To annotate the similarity between primary and reported outcomes, we took pairs of sentences from the corpora annotated for outcomes: the first sentence in each pair comes from the corpus of primary outcomes, the second sentence comes from the corpus of reported outcomes, and both sentences are from the same article (to ensure that primary and reported outcomes exist in the same document, in order to avoid a too high percentage of dissimilar pairs in the final corpus). We used a binary flag to annotate the pairs of outcomes: if both outcomes in a pair are considered to refer to the same outcome, the pair is assigned the 'similar' label; otherwise the 'dissimilar' label. Interestingly, outcomes can refer to the same concept by using antonyms: e.g. "*ICP (Intracranial Pressure) control*" vs. "*uncontrollable intracranial pressure*".

It is important to note that the annotated primary outcomes included all the possible information items present in the sentence (time points, measurement methods, etc.), while the annotated reported outcomes contain the minimal information (usually, the outcome or measurement tool name). Thus, primary outcomes typically contain more information than reported outcomes. When annotating semantic similarity, we disregarded possible differences in additional information such as time points: outcomes were annotated as similar if the outcome/measurement tool used is the same. Table 3 shows some examples of the outcome pairs that were judged to refer to the same (similarity = 1) or different (similarity = 0) concept.

Differences in additional information items (time points, analysis metrics, etc.) are important for a more fine-grained assessment of outcome similarity. However, annotating this information would make the annotation much more complex. We regard comparing additional information on outcomes as a separate task and thus do not include it in the current approach.

Absence of medical knowledge can cause difficulties in annotating outcome similarity. In cases of doubt, the annotator referred to the whole article text or conducted additional research to make the final decision. The total of 3,043 pairs of outcomes were annotated: 701 (612 after deduplication) "similar" and 2,342 (2,187 after deduplication) "dissimilar" pairs.

4.3. Expanded dataset

The ways of referring to an outcome may differ: e.g., the outcome defined as "*the quality of life of people with dementia, as assessed by QoL-AD*" may be referred to by the outcome name ("*the quality of life of people with dementia*") or by the measurement tool name ("*QoL-AD*"), which can in turn be used in the abbreviated or full ("*Quality of Life-Alzheimer's Disease*") form. We expect the variability in choosing one of these options to negatively affect the performance of the similarity

Table 3
Examples of outcomes that are judged as similar(similarity = 1)/different (similarity = 0).

Primary outcome	Reported outcome	Similarity
The change relative to baseline in the multiple sclerosis functional composite score (MSFC)	MSFC score	1
The recruitment rate	The overall recruitment yield	1
The maximum % fall in FEV1 7 h after the first AMP challenge	FEV1	1
ICP control	Uncontrollable intracranial pressure	1
Body weight	Body composition	0
The volume of blood loss between T1 and T4	Bleeding duration	0
Tube dependency at one-year	Hospital admission days	0
HbA1c	Attendance at yoga classes	0

assessment. Thus, we tried to account for this variability in two ways.

First, we searched for abbreviations and their expansions in the full text of the article where a given outcome occurs, using regular expressions. We chose this approach instead of using medical thesauri and automated tools such as Metamap [26] based on the thesauri, because abbreviations can have several possible expansions depending on the particular medical domain. Thus, selecting the correct expansion from a thesaurus would require some additional steps such as detecting the topic of the article. On the contrary, in the text of an article abbreviation expansions are unambiguous. After extracting abbreviations and their expansions, we replace the abbreviations in the outcome mentions by their expansions. For example, for the outcome "*EBM knowledge*" we obtain the expanded variant "*evidence-based medicine knowledge*".

Second, we looked for measurement tool names within outcome mentions, using linguistic markers such as *"measured by"*. We keep the text fragment preceding such markers as the outcome name, and the text following them as the measurement tool name, e.g. for the outcome *"cognitive functioning, as measured by the ADAS-Cog, a 0–70 point scale with a higher score indicating worse cognition"*, we add two variants: *"cognitive functioning"* and *"the ADAS-Cog, a 0–70 point scale with a higher score indicating worse cognition"*.

By applying these algorithms, we obtain an expanded version of the corpus which contains 5,050 pairs of outcomes (1,222 similar and 3,828 dissimilar pairs).

5. Methods

Many existing approaches to semantic similarity assessment rely on manually curated domain-specific resources, such as ontologies or other lexical resources. Although this kind of approach can show good results, its disadvantage consists in the limited word coverage of existing resources and in the need to use tools such as Metamap to map a text to biomedical concepts, resulting in a complex multi-step system with many dependencies.

5.1. Deep learning approach

In the general domain, it was recently shown that unsupervised pre-training of language models on a large corpus, followed by fine-tuning of the models for a particular task, improves the performance of many NLP algorithms, including semantic similarity assessment [43,44]. In the biomedical domain, Blagec and colleagues [38] showed that neural embedding models trained on large domain-specific data outperform the state-of-the-art approaches for similarity assessment.

We explored these novel methods in order to propose an algorithm for assessment of semantic similarity that does not rely on domain-specific resources such as ontologies and taxonomies. We adopt the approach that was recently introduced by Devlin et al. [44] and has already been shown to be highly performant. It consists in fine-tuning language representations that were pre-trained on large datasets, on a limited amount of task-specific annotated data.

Devlin et al. [44] proposed a new method of pre-training language representations, called BERT (Bidirectional Encoder Representations from Transformers). The principle consists in pre-training language representations with the use of a masked language model (MLM) that randomly masks some of the input tokens, allowing pre-training of a deep bidirectional Transformer on both the left and right context. BERT-based pre-trained models can be easily fine-tuned for a supervised task by adding an additional output layer. For our semantic similarity assessment task, we employ the similar architecture as that used for sentence pair classification by Devlin et al. in BERT [44]: a self-attention mechanism is used to encode a concatenated text pair. The task-specific input is fed to the output layer of BERT model, and the end-to-end fine-tuning of all the model parameters is performed. The details on the implementation can be found in Devlin et al. [44].

BERT models were pre-trained on the joint general-domain corpus of English Wikipedia and BooksCorpus, with the total of 3.3B tokens. Two domain-specific version of BERT are of interest for our task: BioBERT [45], pre-trained on a large biomedical corpus of PubMed abstracts and PMC full-text articles comprising 18B tokens, added to the initial BERT training data; and SciBERT [46], pre-trained on a corpus of scientific texts with the total of 3.1B tokens, in addition to the initial BERT training corpus.

BERT provides several models: cased and uncased (differing with regard to the input data preprocessing); base and large (differing with regard to the model size). We fine-tuned and tested both cased and uncased base models. We did not perform experiments with BERT-Large due to limited computational resources. BioBERT has only cased model, with a few versions with different pre-training data (PubMed abstracts only, PMC full-text articles only, or both). We used the model pre-trained on both datasets. SciBERT provides both cased and uncased models and has two versions of vocabulary: BaseVocab (the initial BERT general-domain vocabulary) and SciVocab (the vocabulary built on the scientific corpus). The uncased model with SciVocab is recommended by the authors, as this models showed the best performance in their experiments. We tested both cased and uncased models with SciVocab.

The hyperparameters used for fine-tuning of BERT-based models are shown in the Table 4.

5.2. Baseline approach

We compare the BERT-based approaches to a few simple domain-independent baseline measures that fall into the following categories:

1. string measures:
   - normalized Levenshtein distance [47] (in Tables referred to as *levenshtein_norm*) - the minimal edit distance between two strings (number of edits needed to change one string into the other). We calculate the Levenshtein distance using the Python Levenshtein package and normalize it by dividing it by the length of the longer string.
   - a measure based on the Ratcliff and Obershelp algorithm [48] (in Tables referred to as *difflib*) which calculates the number of matching characters in two strings divided by the total number of characters. We use the implementation proposed by the Python

Table 4
BERT/BioBERT/SciBERT hyperparameters.

Hyperparameter	Value	Definition
do_lower_case	True (uncased models)/False (cased models)	Whether to lower case the input text
max_seq_length	128	The maximum total input sequence length after WordPiece tokenization
train_batch_size	32	Total batch size for training
eval_batch_size	8	Total batch size for eval
predict_batch_size	8	Total batch size for predict
learning_rate	5e-5	The initial learning rate for Adam
num_train_epochs	3.0	Total number of training epochs to perform
warmup _proportion	0.1	Proportion of training to perform linear learning rate warmup for
save_checkpoints _steps	1000	How often to save the model checkpoint
iterations_per _loop	1000	How many steps to make in each estimator call
use_tpu	False	Whether to use TPU or GPU/CPU
master	None	TensorFlow master URL

difflib library (SequenceMatcher function).

2. lexical measures reflecting the number of lexical items shared by the compared phrases:
   - the proportion of lemmas occurring in both compared outcomes (in Tables referred to as *lemmas*), calculated as the proportion of the lemmas shared by the compared phrases divided by the length (in lemmas) of the shorter outcome. Lemmatization was performed with the help of WordNetLemmatizer function of Python NLTK library.
   - the proportion of stems occurring in both compared outcomes (in Tables referred to as *stems*), calculated as the proportion of the stems shared by the compared phrases divided by the length (in stems) of the shorter outcome. Stemming was performed using the PorterStemmer function of Python NLTK library.

   In both lexical measures, stop-words and digits were excluded, as well as some words with general semantics typical for outcome mentions (e.g. “change”, “increase”, “difference”).
3. vector-based measures:
   - a cosine similarity between the compared outcomes (in Tables referred to as *gensim*), using vector representation obtained with Latent Semantic Analysis using singular value decomposition. We use the implementation proposed by the Python gensim [49] library[8].
   - a cosine similarity between the compared outcomes (in Tables referred to as *spacy*), using an average of word vectors. We use the implementation proposed by the Python spaCy [50] library.
4. ontology-based measures:
   - path similarity score (in Tables referred to as *path*) is a WordNet-based measure of similarity of two word senses calculated as the shortest path connecting them in the hypernym–hyponym taxonomy.
   - Leacock-Chodorow similarity score [19] (in Tables referred to as *lch*) is a WordNet-based measure of similarity of two word senses based on the shortest path connecting them and the maximum depth of the taxonomy in which they are found.
   - Wu-Palmer similarity score [20] (in Tables referred to as *wup*) is a WordNet-based measure of similarity of two word senses based on the depth of the senses in the taxonomy and that of their most specific ancestor node.

   For all three measures, we use the functions implemented in the Python NLTK library. The final scores are calculated as proposed by Mihalcea and colleagues [51].

Each of these measures returns a similarity score on a certain scale (most typically, between 0 and 1). After testing several cut-off values, we manually set a threshold for each measure to maximize the F-measure: pairs of outcomes with the similarity measure above the threshold are considered similar. The thresholds chosen for each measure are shown in Table 5.

5.3. Feature-based machine-learning approach

Following the approach proposed by Soganciogl u et al. [18], we trained and tested a number of classifiers, taking the above-listed single similarity measures as the input features. We evaluated several classifiers: Support Vector Machine (SVM) [52]; Decision Tree Classifier [53]; MLP Classifier [54]; K-neighbors Classifier [55]; Gaussian Process Classifier [56]; Random Forest Classifier [57]; Ada Boost Classifier [58]; Extra Trees Classifier [59]; Gradient Boosting Classifier [60]. We used the implementation provided by Python scikit-learn library [61]. We performed hyperparameters tuning via exhaustive grid search (with the help of the scikit-learn GridSearchCV function). The chosen hyperparameters are shown in Table 6 (for the experiments on the original corpus) and Table 7 (for the experiments on the expanded corpus).

5.4. Experiments on the expanded dataset

The expanded dataset (with expanded abbreviations and added variants of referring to an outcome by the measurement tool name or by the outcome name) is used in the experiments in the following way. For individual similarity measures, we compare all the combinations of variants for both outcomes. Out of the similarity scores obtained for all the variants, we take the maximum value as the final evaluation score. For machine learning approaches, we expanded the original annotated corpus by the extracted variants of the outcomes. We trained and tested the machine learning and deep learning approaches on both the original corpus an on the expanded corpus.

6. Results and discussion

For the deep learning approach, we performed the evaluation using 10-fold cross-validation, with the dataset split into train and development sets in the proportion 9:1. The performance is reported for the development set. For scikit-learn classifiers, we performed 10-fold cross-validation using the scikit-learn built-in cross_validate function.

Table 8 below presents the results of our experiments on the original and expanded corpus, respectively. We use the following notations in the results tables: *BioBERT*, *SciBERT uncased*, *SciBERT cased*, *BERT uncased* and *BERT cased* refer to the results of fine-tuning of the corresponding language model. *RandomForest*, *MLP*, *GaussianProcess*, *GradientBoosting*, *KNeighbors*, *ExtraTrees*, *AdaBoost*, *DecisionTree*, and *SVC* refer to the results of the corresponding scikit-learn classifier. *stems* and *lemmas* refer to the lexical similarity measures (the proportion of stems/lemmas occurring in both compared outcomes). *gensim* and *spacy* refer to vector-based measures (cosine similarity as implemented by gensim and spacy packages, respectively). *levenshtein_norm* refers to the normalized Levenshtein distance, *difflib* refers to the Ratcliff and Obershelp

[8] https://radimrehurek.com/gensim/tut3.html.

Table 5
Thresholds set for the similarity measures.

Measure	Threshold
difflib	0.4
levenshtein_norm	0.3
lemmas	0.6
spacy	0.6
gensim	0.9
stems	0.6
path	0.4
wup	0.5
lch	2.5

Table 6
Hyperparameters for classifiers on the original corpus.

Classifier	Hyperparameters
RandomForest	max_depth = 25, min_samples_split = 5, n_estimators = 300
MLP	activation = 'tanh', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate = 'constant', solver = 'adam'
GaussianProcess	1.0 ∗RBF(1.0)
GradientBoosting	default
KNeighbors	n_neighbors = 13, p = 1
ExtraTrees	default
AdaBoost	default
DecisionTree	default
SVC	C = 1000, gamma = 0.001, kernel = 'rbf'

Table 7
Hyperparameters for classifiers on the expanded corpus.

Classifier	Hyperparameters
RandomForest	max_depth = 25, min_samples_split = 5, n_estimators = 300
KNeighbors	n_neighbors = 9, p = 5
GradientBoosting	learning_rate = 0.25, max_depth = 23.0, max_features = 7, min_samples_leaf = 0.1, min_samples_split = 0.2, n_estimators = 200
MLP	activation = 'relu', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate = 'adaptive', solver = 'adam'
GaussianProcess	1.0 ∗RBF(1.0)
ExtraTrees	default
AdaBoost	learning_rate = 0.1, n_estimators = 500
SVC	kernel = 'linear', C = 1, random_state = 0
DecisionTree	max_depth = 1.0, max_features = 2, min_samples_leaf = 0.1, min_samples_split = 1.0

algorithm-based measure. *path* refers to the path similarity score; *lch* refers to the Leacock-Chodorow similarity score; *wup* refers to the Wu-Palmer similarity score.

Among the single similarity measures tested on our original (non-expanded) corpus, the best performance was shown by the stem-based measure (F-measure = 71.35%). Among the classifiers using the combination of measures as features, the best results were achieved by the Random Forest Classifier (F-measure = 84.73%). Among the deep learning models, the fine-tuned BioBERT model showed the highest performance (F-measure = 89.75%).

These results clearly show that, out of the three tested approaches (baseline single similarity measures, machine learning classifiers using the single measures as features, and deep learning), the best results on the original corpus were shown by the deep learning approaches. All the single measures were inferior to the classifiers based on the combination of the single measures. Thus, we can state the measures complement each other. Further, all the deep learning BERT-based approaches showed better performance than each of the classifiers, which indicates that the pre-trained representations are more powerful in reflecting semantic similarity than the measures used.

On the expanded corpus, the performance of single measures changed slightly compared to that on the original corpus (cf. Table 8).

Table 8
Results.

	On the original corpus			On the expanded corpus		
Algorithm	Precision	Recall	F1	Precision	Recall	F1
BioBERT	88.93	90.76	89.75	92.98	93.85	93.38
SciBERT uncased	87.99	90.78	89.3	91.3	91.79	91.51
SciBERT cased	87.31	91.53	89.3	89	92.54	90.69
BERT uncased	85.76	88.15	86.8	89.31	89.12	89.16
RandomForest	86.76	82.92	84.73	74.09	60.12	66.13
BERT cased	83.36	85.2	84.21	88.25	90.1	89.12
MLP	87.79	80.61	83.95	72.21	58.05	63.87
GaussianProcess	86.69	81.11	83.74	72.08	57.13	63.58
GradientBoosting	87.84	79.96	83.63	72.94	58.4	64.72
KNeighbors	87.35	78.81	82.75	75.24	58.13	65.31
ExtraTrees	85.26	79.29	82.08	71.83	57.14	63.47
AdaBoost	86.08	77.99	81.79	72.66	55.87	62.97
DecisionTree	81.66	79.62	80.53	62.73	63.09	60.61
SVC	82.3	78.32	80.19	73.2	54.42	62.26
stems	64.03	80.56	71.35	64.03	80.56	71.35
lemmas	64.75	77.45	70.54	63.18	78.23	69.91
gensim	55.71	83.66	66.88	54.98	79.14	64.89
path	60.06	65.36	62.6	58.04	69.47	63.24
wup	53.26	68.14	59.78	52.15	73.35	60.96
levenshtein_norm	65.87	49.84	56.74	64.64	56.14	60.09
difflib	47.08	71.08	56.64	63.84	61.73	62.77
lch	59.42	53.59	56.36	62.95	25.02	35.81
spacy	35.86	75.65	48.66	35.86	75.65	48.66

The best result, achieved by the stem-based measure, was not improved. The performance of machine learning classifiers on the expanded corpus dropped significantly (the highest F-measure was 66.13% vs. 84.73% on the original corpus). On the contrary, the performance of all the fine-tuned deep learning models was better on the expanded corpus than on the original corpus. The best result, similarly to the original corpus, was shown by the fine-tuned BioBERT model: F-measure was 93.38%.

6.1. Error analysis

We provide here the error analysis of the best-performing model (fine-tuned BioBERT) on the original corpus. The most common cases of errors are as follows:

1. Use of abbreviations which leads to false negatives, e.g.:
   - *Uncontrollable intracranial pressure – **ICP** control*
   - *sickness absence – **SA** days*
   - *pain catastrophising – global **PC***
   - *controlling intracranial pressure – **ICP** control*
   - *the Yale-Brown Obsessive–Compulsive Scale – the change in **YBOCS** score from baseline to endpoint*
2. Terms that are semantically close but refer to different measured variables result in false positives, e.g.:
   - *coma **recovery** time – total coma **duration***
   - *patient **satisfaction** – patient **comfort***
   - *time to **azoospermia** time to severe **oligozoospermia***

   In particular, this type of error can be observed when the terms are hyponyms of the same term, e.g.:
   - ***child*** body mass index (BMI) z-score – ***parent*** BMI
   - ***foot*** pain – ***'first-step'*** pain
   - *the proportion of delivered compressions **within** target depth compared over a 2-min period within the groups and between the groups – the proportion of delivered compressions **below** target depth*

   Besides, this type of error occurs when the outcomes refer to different aspects of one parameter, e.g. (words indicating the differences in semantics of the phrases are in bold):
   - *the GSRS **subscores** for abdominal pain – the GSRS **total score***
   - *The **frequency** of acute exacerbation – **duration** of acute*

exacerbation

- ***costs*** per quality adjusted life years (cost/QALY) – Quality adjusted life years
- ***time*** needed to perform the motor task – ***degree of help*** needed to perform the task
- *the mean time to onset of the **first 24-h** heartburn-free period after initial dosing – The mean number of heartburn-free days **by D7***
- *the proportion of patients with plasma HIV-1 RNA levels **<200** copies/mL at week 24 – HIV-1 RNA **<50** copies/mL*

3. Use of terms for which the similarity can only be established based on domain knowledge but not by their textual features leads to false negatives, e.g.:
   - *HSCL-25 – the severity of symptoms of depression and anxiety* (HSCL-25 is a checklist measuring the symptoms of anxiety and depression[9])
   - *response rate – took part in the Link-Up Study*
   - *return of final follow-up questionnaire or reminder by the participant – the response rates*
4. Significantly different level of detail in two mentions of the same measure can lead to false negatives, e.g.:
   - *the incidence of oxygen desaturations defined as a decrease in oxygen saturation ⩾5%, assessed by continuous pulse oxymetry, at any time between the start of the induction sequence and two minutes after the completion of the intubation – oxygen desaturations*

6.2. The best method for assessing semantic similarity

On the original outcome pairs corpus, the best-performing single similarity measure is the stem-based one (F1 = 71.35%), followed by the lemmas-based and gensim measures (Table 8). The gensim measure shows the best recall (83.66%).

All the scikit-learn classifiers trained on the original corpus using the combination of the single measures as features outperformed single measures (Table 8). The best results were achieved by the Random Forest Classifier (F-measure of 84.73%).

When trained on the original corpus, all the BERT-based models, except for the one using the BERT cased model, outperformed the feature-based classifiers and single similarity measures (Table 8). The best results were shown by the fine-tuned BioBERT model, reaching the F-measure of 89.75%. Results of fine-tuned SciBERT models (both cased and uncased) reached the F-measure of 89.3%, closely following BioBERT; the SciBERT cased model demonstrated the best recall (91.53%).

These results show that fine-tuned models using deep pre-trained language representations can outperform all the other tested similarity measures, with an additional advantage of not requiring any specialized resources or specific text preprocessing such as mapping to the UMLS concepts. Pre-training of language models on biomedical texts proves to be an advantageous approach as it allows to learn representations for domain-specific words, including abbreviations, from the available large unstructured data.

6.3. Does the addition of variants of referring to an outcome help?

For the single measures of similarity, expansion of the corpus by the variants of outcomes improved the performance of Wordnet-based and string-based measures, but did not improve the results of the three best-performing measures - stem- and lemma-based ones and the gensim measure (cf. Table 8).

A possible explanation for the absence of improvement in the stem- and lemma-based measures is that the primary outcomes are usually rather lengthy and detailed, and tend to include all the variants: abbreviations and their expansions, measurement method. Thus, additional variants are not in fact needed. For example, the primary outcome "*depression severity measured by the Beck Depression Inventory-II (BDI-II)*" may be reported as "*depression severity*", "*the Beck Depression Inventory-II*" or "*BDI-II*", but all these variants are already present within the primary outcome mention, thus, measuring the intersection in terms of stems or lemmas will return a high similarity score. At the same time, for string-based and WordNet-based measures, addition of variants is useful: for the example above, if the outcome is reported as "*BDI-II*", it will be expanded as "*the Beck Depression Inventory-II*", which will have high similarity scores with the variant "*the Beck Depression Inventory-II (BDI-II)*" of the primary outcome.

For the classifiers using single similarity measures as features, adding outcome variants to the training corpus did not prove useful: the results of the classifiers trained on the corpus expanded by the outcome variants dropped significantly (cf. Table 8).

It should be highlighted that single measures and classifiers in our approach account for outcome variants in different ways: single measures compare all the pairs of variants and take the highest score as the final result, thus, low similarity between some of the variants does not affect the results. On the contrary, the classifiers use the expanded corpus to train, and thus, pairs of variants with low similarity scores but with the 'similar' label can negatively impact the results.

Interestingly, the addition of the variants to the training corpus can be useful: performance of all the BERT-based systems improved on the corpus expanded by outcome variants (cf. Table 8). The best result was achieved by the fine-tuned BioBERT model, with the F-measure of 93.38%.

The difference between the results of classifiers using single measures as features and the fine-tuned BERT-based models on the expanded corpus demonstrates differences between these approaches. BERT-based models successfully train on the expanded corpus as they use deep pre-trained language representations and fine-tune to learn the features required for a given task, while the training of classifiers is likely to be undermined by the pairs of outcome variants with low scores on the single similarity measures.

The results of these experiments should, however, be taken with caution, as the expansion of the corpus by outcome variants was performed automatically. We manually checked the quality of the algorithm, but it does not exclude presence of some noise. Still, we believe that this approach is promising for our task.

6.4. What metrics are best able to identify similar or dissimilar outcomes?

Out of single similarity measures, the best ability to distinguish between similar and dissimilar outcomes, in both the original and the expanded corpora, was shown by the stem-based measure, followed by the lemma-based measure (Table 8).

6.5. What classifiers are best able to distinguish between similar and dissimilar outcome pairs?

In our experiments, the Random Forest Classifier showed the best results in the task of distinguishing between similar and dissimilar outcome pairs, compared to a range of other classifiers (MLP, Gaussian Process Classifier, Gradient Boosting Classifier, K-neighbors Classifier, Extra Trees Classifier, Ada Boost Classifier, Decision Tree Classifier, and SVM) (Table 8).

6.6. What language representation is best able to represent outcomes?

Our experiments showed that the best performance for semantic similarity assessment of outcomes is shown by the fine-tuned BioBERT model, i.e. a language model pre-trained on a large (18B tokens) biomedical corpus in addition to a 3.3.B tokens general domain-corpus. This model outperformed the models trained on the general-domain corpus only (BERT) and the models trained on a smaller (3.1) corpus of scientific paper in addition to the general domain corpus (SciBERT)

[9] http://hprt-cambridge.org/screening/hopkins-symptom-checklist/.

(Tables 8).

7. Conclusion

Evaluation of similarity assessment of trial outcomes is a vital part of tasks such as assessment of an article for outcome switching, reporting bias and spin; besides, it can be used to improve the adherence to Core Outcomes Sets use. In this work, we introduced a first open-access corpus of pairs of primary and reported outcomes, annotated on a binary scale as similar or different. We presented our experiments on developing an algorithm of semantic similarity assessment not using domain-specific resources such as ontologies and taxonomies. We tested a number of single similarity measures, classifiers using the combination of single measures as features, and a number of deep learning models. We explored the possibility of using variants of referring to outcomes (abbreviations, measurement tool names) to improve the performance of similarity assessment.

The best results were shown by the deep learning approach using the BioBERT fine-tuned model, both on the original corpus and on the corpus expanded by the outcome variants.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgements

We thank prof. Isabelle Boutron from the University Paris Descartes, prof. Patrick Bossuyt from the University of Amsterdam, and Dr. Liz Wager from SideView who provided valuable insight and expertise as our consultants in the domain of reporting of clinical trials.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 676207.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.yjbinx.2019.100058.

References

[1] P. Smith, R. Morrow, D. Ross, Outcome measures and case definition, Field Trials of Health Interventions: A Toolbox, 3rd ed., OUP Oxford, 2015.
[2] M. Ghert, The reporting of outcomes in randomised controlled trials: The switch and the spin, Bone Joint Res. 6 (2017) 600–601, https://doi.org/10.1302/2046-3758.610.BJR-2017-0296.
[3] B. Goldacre, H. Drysdale, A. Powell-Smith, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, K. Mahtani, C. Heneghan, The compare trials project, 2016. URL www.COMPare-trials.org.
[4] B. Goldacre, H. Drysdale, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, A. Powell-Smith, C. Heneghan, K.R. Mahtani, Compare: a prospective cohort study correcting and monitoring 58 misreported trials in real time, Trials 20 (1) (2019) 118, https://doi.org/10.1186/s13063-019-3173-2.
[5] E. Slade, H. Drysdale, B. Goldacre, Discrepancies between prespecified and reported outcomes, BMJ (2015), <http://www.bmj.com/content/351/bmj.h5627/rr-12>.
[6] J. Weston, K. Dwan, D. Altman, M. Clarke, C. Gamble, S. Schroter, P. Williamson, J. Kirkham, Feasibility study to examine discrepancy rates in prespecified and reported outcomes in articles submitted to the bmj, BMJ Open (2016), https://doi.org/10.1136/bmjopen-2015-010075 <http://bmjopen.bmj.com/content/6/4/e010075>.
[7] D. Altman, D. Moher, K. Schulz, Harms of outcome switching in reports of randomised trials: Consort perspective, BMJ: British Med. J. (Online) (2017).
[8] I. Boutron, S. Dutton, P. Ravaud, D. Altman, Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes, JAMA 303 (20) (2010) 2058–2064, https://doi.org/10.1001/jama.2010.651.
[9] S. Lockyer, R. Hodgson, J. Dumville, N. Cullum, spin in wound care research: The reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes, Trials 14 (2013) 371, https://doi.org/10.1186/1745-6215-14-371.
[10] C. Lazarus, R. Haneef, P. Ravaud, I. Boutron, Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention, BMC Med. Res. Methodol. 15 (1) (2015) 85, https://doi.org/10.1186/s12874-015-0079-x.
[11] K. Chiu, Q. Grundy, L. Bero, 'spin' in published biomedical literature: A methodological systematic review, PLOS Biol. 15 (2017) e2002173, https://doi.org/10.1371/journal.pbio.2002173.
[12] J. Diong, A.A. Butler, S.C. Gandevia, M.E. Héroux, Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice, PLoS One (2018), https://doi.org/10.1371/journal.pone.0202121.
[13] I. Boutron, P. Ravaud, Misrepresentation and distortion of research in biomedical literature, Proc. Natl. Acad. Sci. U S A 115 (11) (2018) 2613–2619, https://doi.org/10.1073/pnas.1710755115.
[14] I. Boutron, D. Altman, S. Hopewell, F. Vera-Badillo, I. Tannock, P. Ravaud, Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the spiin randomized controlled trial, J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol. 32 (11) (2014), https://doi.org/10.1200/JCO.2014.56.7503.
[15] R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, I. Boutron, Interpretation of results of studies evaluating an intervention highlighted in google health news: A cross-sectional study of news, PloS One 10 (2015) e0140889, https://doi.org/10.1371/journal.pone.0140889.
[16] A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, P. Ravaud, Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study, PLOS Med. 9 (9) (2012) 1–11, https://doi.org/10.1371/journal.pmed.1001308.
[17] T. Pedersen, S.V. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. 40 (3) (2007) 288–299, https://doi.org/10.1016/j.jbi.2006.06.004 <http://www.sciencedirect.com/science/article/pii/S1532046406000645>.
[18] G. Soganciogl u, H. Öztürk, A. Özgür, Biosses: a semantic sentence similarity estimation system for the biomedical domain, Bioinformatics 33 (2017) 14, https://doi.org/10.1093/bioinformatics/btx238.
[19] C. Leacock, M. Chodorow, Combining Local Context and WordNet Similarity for Word Sense Identification, vol. 49, MITP, 1998, pp. 265–.
[20] Z. Wu, M. Palmer, Verbs semantics and lexical selection, Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA, 1994, pp. 133–138, , https://doi.org/10.3115/981732.981751.
[21] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Trans. Systems, Man, Cybernet. 19 (1989) 17–30.
[22] J.E. Caviedes, J.J. Cimino, Towards the development of a conceptual distance metric for the umls, J. Biomed. Inform. 37 (2) (2004) 77–85, https://doi.org/10.1016/j.jbi.2004.02.001 <http://www.sciencedirect.com/science/article/pii/S1532046404000218>.
[23] C. Fellbaum, WordNet: An electronic lexical database (Language, Speech, and Communication), The MIT Press, Cambridge, MA, 1998.
[24] B.T. McInnes, T. Pedersen, S.V.S. Pakhomov, Umls-interface and umls-similarity: Open source software for measuring paths and semantic similarity, AMIA Annual Symposium proceedings. AMIA Symposium 2009, 2009, pp. 431–435.
[25] P. Lord, R. Stevens, A. Brass, C. Goble, Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation, Bioinformatics 19 (2003) 1275–1283.
[26] A. Aronson, Effective mapping of biomedical text to the umls metathesaurus: The metamap program, in: AMIA Annual Symposium 2001, 2001, pp. 17–21.
[27] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 448–453 <http://dl.acm.org/citation.cfm?id=1625855.1625914>.
[28] D. Lin, An information-theoretic definition of similarity, Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 296–304 <http://dl.acm.org/citation.cfm?id=645527.657297>.
[29] D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective, J. Biomed. Inform. 44 (5) (2011) 749–759, https://doi.org/10.1016/j.jbi.2011.03.013 <http://www.sciencedirect.com/science/article/pii/S1532046411000645>.
[30] M.B. Aouicha, M.A.H. Taieb, Computing semantic similarity between biomedical concepts using new information content approach, J. Biomed. Inform. 59 (2016) 258–275, https://doi.org/10.1016/j.jbi.2015.12.007 <http://www.sciencedirect.com/science/article/pii/S1532046415002877>.
[31] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, J. Montmain, A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain, J. Biomed. Inform. 48 (2014) 38–53, https://doi.org/10.1016/j.jbi.2013.11.006 <http://www.sciencedirect.com/science/article/pii/S1532046413001834>.
[32] I. Spasić, S. Ananiadou, A flexible measure of contextual similarity for biomedical terms, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2004, pp. 197–208.
[33] W. Blacoe, M. Lapata, A comparison of vector-based representations for semantic composition, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 546–556. https://www.aclweb.org/anthology/D12-1050.
[34] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., USA, 2013, pp. 3111–3119. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.
[35] S. Henry, C. Cuffy, B.T. McInnes, Vector representations of multi-word terms for semantic relatedness, J. Biomed. Inform. 77 (2018) 111–119, https://doi.org/10.

1016/j.jbi.2017.12.006 <http://www.sciencedirect.com/science/article/pii/S1532046417302769>.
[36] J. Park, K. Kim, W. Hwang, D. Lee, Concept embedding to measure semantic relatedness for biomedical information ontologies, J. Biomed. Inform. 94 (2019) 103182, https://doi.org/10.1016/j.jbi.2019.103182 <http://www.sciencedirect.com/science/article/pii/S1532046419301005>.
[37] S. Henry, A. McQuilkin, B.T. McInnes, Association measures for estimating semantic similarity and relatedness between biomedical concepts, Artif. Intell. Med. 93 (2019) 1–10, https://doi.org/10.1016/j.artmed.2018.08.006 extracting and Processing of Rich Semantics from Medical Texts. <http://www.sciencedirect.com/science/article/pii/S0933365717304475>.
[38] K. Blagec, H. Xu, A. Agibetov, M. Samwald, Neural sentence embedding models for semantic similarity estimation in the biomedical domain, BMC Bioinformat. (2019) 178.
[39] S.V.S. Pakhomov, T. Pedersen, B. Mcinnes, G.B. Melton, A. Ruggieri, C. Chute, Towards a framework for developing semantic relatedness reference standards, J. Biomed. Informat. 44 (2010) 251–265, https://doi.org/10.1016/j.jbi.2010.10.004.
[40] S. Pakhomov, B. Mcinnes, T. Adam, Y. Liu, T. Pedersen, G.B. Melton, Semantic similarity and relatedness between clinical terms: An experimental study, AMIA, Annual Symposium Proceedings/ AMIA Symposium. AMIA Symposium 2010, 2010, pp. 572–576.
[41] Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, H. Liu, Medsts: A resource for clinical semantic textual similarity, CoRR abs/1808.09397, 2018. arXiv:1808.09397. URL http://arxiv.org/abs/1808.09397.
[42] A. Koroleva, Annotated corpus for semantic similarity of clinical trial outcomes, May 2019. https://doi.org/10.5281/zenodo.3234827.
[43] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding with unsupervised learning, Technical report, OpenAI, 2018.
[44] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805, 2018. arXiv:1810.04805. URL http://arxiv.org/abs/1810.04805.
[45] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, arXiv preprint arXiv:1901.08746, 2019.
[46] I. Beltagy, A. Cohan, K. Lo, Scibert: Pretrained contextualized embeddings for scientific text, arXiv preprint arXiv:1903.10676, 2019.
[47] F.P. Miller, A.F. Vandome, J. McBrewster, Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance, Alpha Press, 2009.
[48] J. Ratcliff, D. Metzener, Pattern matching: The gestalt approach, Dr. Dobb's J. (Jul 1998).
[49] R. Rehurek, P. Sojka, Software framework for topic modelling with large corpora, Proc. LREC Workshop on New Challenges for NLP Frameworks, 2010, pp. 2216–2219.
[50] M. Honnibal, M. Johnson, An improved non-monotonic transition system for dependency parsing, in: Proc. of EMNLP 2015, ACL, Lisbon, Portugal, 2015, pp. 1373–1378. https://aclweb.org/anthology/D/D15/D15-1162.
[51] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, in: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06, AAAI Press, 2006, pp. 775–780. URL http://dl.acm.org/citation.cfm?id=1597538.1597662.
[52] C. Cortes, V. Vapnik, Support-vector networks, Machine Learn. 20 (3) (1995) 273–297, https://doi.org/10.1007/BF00994018.
[53] L. Rokach, O. Maimon, Data Mining with Decision Trees: Theroy and Applications, World Scientific Publishing Co. Inc., River Edge, NJ, USA, 2008.
[54] C. von der Malsburg, Frank Rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms, Brain Theory (1986) 245–248, https://doi.org/10.1007/978-3-642-70911-1_20.
[55] N. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Statist.- AMER STATIST 46 (1992) 175–185, https://doi.org/10.1080/00031305.1992.10475879.
[56] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.
[57] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[58] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139, https://doi.org/10.1006/jcss.1997.1504 <http://www.sciencedirect.com/science/article/pii/S002200009791504X>.
[59] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.
[60] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2.
[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Machine Learn. Res. 12 (2011) 2825–2830.