Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SMAFIRA-c: A benchmark text corpus for evaluation of approaches to relevance ranking and knowledge discovery in the biomedical domain

Daniel Butzke (Daniel.Butzke@bfr.bund.de)
Bundesinstitut fur Risikobewertung https://orcid.org/0000-0002-4800-4655
Nadine Dulisch
GESIS Leibniz-Institut fur Sozialwissenschaften in Koln
Sebastian Dunst
Bundesinstitut fur Risikobewertung
Matthias Steinfath
Bundesinstitut fur Risikobewertung
Mariana Neves
Bundesinstitut fur Risikobewertung
Brigitte Mathiak
GESIS Leibniz-Institut fur Sozialwissenschaften in Koln
Barbara Grune
Bundesinstitut fur Risikobewertung

Research article

Keywords: Relevance ranking, knowledge discovery, case study, biomedicine, animal use alternatives

Posted Date: March 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-16454/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Loading [MathJax]/jax/output/CommonHTML/jax.js

Abstract

Background The engineering of elaborate and innovative tools to navigate the ever growing biomedical knowledge base, instanced in PubMed/Medline, must be guided by genuine case studies addressing *realw* or *ld´user ≠ eds. Furtherm* or *e, alg* or *ithm - basedpredictionsregard ∈ g*similarity´, *related ≠ ss´* or relevance´ of pieces of information (e.g. relevance ranking) should be transparent and comprehensible to users.

Results We here present a corpus of abstracts (n = 300) annotated on document level representing three case studies in the experimental biomedical domain. The SMAFIRA corpus mirrors *real - w* or *ld´ ∈ f* or *mationretrieval ≠ eds, i. e. thent* if *icationofpotentiala < ernatives → givenanimal*exp*erimentst* $\hat{\supset}$ *p* or *t*equivalent´ scientific purposes while using basically different experimental methodology. Since in most cases not even the authors of *re ≤ vant´researchpapersareawareofsuchapossib ≤ implicationoftheir*exp*erimentalapproaches, ourcasestudiesactuallyillustrateknow ≤ d* equivalence´) was conducted by one researcher with broad domain knowledge (in one case study supported by a second opinion from a domain expert) and was informed by a newly created model describing distinguishable stages in experimental biomedicine. Furthermore, such stages were linked to generic scientific purposes. This perspective thus may share some commonalities with topic modelling approaches. Annotation of *re ≤ vance´* (*i. e.* equivalence´ of scientific purpose plus alternative methodology) relied on expert knowledge in the domain of animal use alternatives. The case studies were used for an evaluation of rankings which were provided by the `similar articles´ algorithm employed in PubMed.

Conclusions Building on approved techniques utilized in the domain of intellectual property, we have adapted the concept of `equivalence´ to support a transparent, reproducible and stringent comparison of biomedical textual documents with regards to the implied scientific objectives. This concept may allow for text mining with improved resolution and may aid the retrieval of appropriate animal use alternatives. Computer science researchers in the field of biomedical knowledge discovery may also use our corpus, which is designed to grow essentially in the near future, as a reliable and informative benchmark for the evaluation of algorithms supporting such a goal. Annotations are available from GitHub.

Background

In the biosciences, information retrieval (IR) is as important as DNA-sequencing, protein biochemistry or cell imaging, since all new experimental findings must be interpreted in light of the existing biomedical knowledge base.

A prominent resource for biomedical text-based IR is PubMed/MEDLINE. Currently, PubMed provides access to more than 29 million citations for biomedical literature. Besides keyword-based queries, PubMed supports publication *similarity´ - basedqueries. Theretrievalofsuch*similar articles´ is fueled by the pmra-algorithm (1) which considers *contentsimilarity´, i. e.* similarity´ "in terms of the topics or concepts that they are about". The number of

similarartic ≤ s´assig ≠ d → sing ≤ citations ∈ PubMedcanvaryomsomedozens → sometensofthous and *s.* Sin*cem* or *ethan*80 % *ofPubMe*

similar articles´) there is a need for ranking. Thus, *similarartic ≤ s´arerankeda* or *d ∈ g → their*similarity score´ with regards to the reference publication, from highest to lowest. The most *re ≤ vant´ ∈ f* or *mation → agivenretrieval ≠ ed, however, µst¬ ≠ cessarilybe ∈ cluded ∈ sucha* ⊤ *- 20col ≤ ction, butmaybepositio ≠ da* equivalent´ scientific purposes while using basically different experimental methodology (animal use alternatives, see below) - a *similarartic ≤ s´searchisonly∂ lyhelpf,* sin*cethemost*similar´ research to any animal experiment, as judged by pmra, always is another animal experiment (with `similar´ scientific objectives).

Nevertheless, pmra does principally retrieve information too which is *re ≤ vant´ ∈ termsofanimalusea < ernatives. Butsuch ∈ f* or *mationmayberankedatpositions, whichare → odis*tan*tomthe* ⊤ *- 20 → becons* animal use alternatives´, on PubMed-similar-articles-corpora is possible, but – unfortunately – it requires that a priori annotation with respective terms had been conducted by MEDLINE indexing personnel already. To assign such terms, however, the indexers (or supporting tools like NLM Medical Text Indexer (3)) rely on clear indications that a given piece of information belongs to the class *animalusea < ernatives´. Suchac ≤ ar ∈ dicationcodbeapublication ∈ aspecialjournallikeATLA*(*A < ernatives → Lab* or *a → ryAnimals*)*f* animal use alternatives´-cluster. In fact, when inspecting the MeSH-terms assigned to the publications judged *re ≤ vant´ ∈ thisstudy*(*n = 33*), *wefound, tĥo ≠ waslabe ≤ das*animal use alternative´.

Information retrieval as legal act

Researchers who plan to perform a scientific project involving animal experimentation in one of the Member States of the European Union have to file an application for project authorization (Article 37, Directive 2010/63/EU (4)). This application shall include information (specified in Annex VI, Directive 2010/63/EU (4)) which must be gathered via an evaluation of the current scientific knowledge. The legally required information includes 1.) the relevance and justification of the use of animals, 2.) the application of methods to "replace, reduce and refine" the use of animals in procedures (3R principle(5)), and 3.) the avoidance of unjustified duplication of procedures. Since PubMed/MEDLINE is a prominent resource representing the current biomedical knowledge, researchers routinely use this resource (and its search tools) to fulfill these legal requirements. An appealing approach with apparent ease is to employ the *similarartic ≤ s´ → olderıctedabove: justıınp∮ arelatedpublicationdescrib ∈ g*similar´ in vivo research and then screen the *similarartic ≤ s´col ≤ ctionf* or relevant´ abstracts. While the pmra-algorithm may well be suited to address information need number three ("avoidance

Loading [MathJax]/jax/output/CommonHTML/jax.js

of duplication") by helping to retrieve `similar´ research, it is not yet optimized to help users with retrieving publications about methods to replace the use of animals (i.e. information need number two).

Introducing the SMAFIRA project

We have initiated a project that aims to support a 3R relevant IR. The SMAFIRA project (smart feature based interactive re-ranking) so far has resulted in the completion of a tool for the preparation of annotated test sets (case studies in the domain of biomedicine) and the assessment of algorithms implemented in the WEKA library (6) using these case studies. Annotations for seven case studies were assigned on document level and comprised judgements on $\equiv a \leq nce´$, relevance´ and $animaluse´$. Equivalence´ regards accordance in scientific objective(s) and comparability of experimental results, and thereby consequently ignores methodology (animal experimentation in particular).
$Re \leq vance´ consrsthepossib \leq impactoftheusedmethodology withregard \rightarrow the3Rpr \in cip \leq ,$ and animal use´ regards the kind of animal use deducible from the abstracts of citations (e.g. in vivo AND/OR ex vivo).
$Re \leq vance´ was\det er \min edbaseduponthestipationsofDirective\frac{2010}{63}/EU: thus, any$relevant´ experimental approach would be appropriate to replace the use of live vertebrate animals (or cephalopods). Research using live invertebrate animals (e.g. flies) instead of mice would be deemed `relevant´ as well, although such research indeed is undertaken in vivo, since flies are not protected under the animal protection law.

Box 1: Glossary

`Equivalence´	**Overlap in critical scientific entities with regard to the scientific objectives of two publications under comparison (completely or partly). The critical scientific entities for comparison are itemized in a chart.**
`Relevance´	Experimental approaches that would be appropriate to replace the use of live vertebrate animals (or cephalopods) in a given research project with defined scientific objective. Subset of the 3R principle.

The basic problem in setting up such sets of annotated test publications for evaluation of algorithms is boiled down to an essence in (7): "Evaluating the performance of … algorithms is a challenging task. It is challenging not only because manually created gold standards are required, but also because creating such gold standards is not a well-defined task."

With SMAFIRA, we therefore attempted to render the task of our gold-standard-creation (regarding the annotated label $\equiv a \leq nce´) as\mu ch$well-defined´ as possible and built on a technique already accredited in another context:

In the domain of intellectual property the infringement of a patent can be considered with the aid of an $\in \in \geq mentanalysis´. Thee \leq mentsofthepatent'sclaimsarelisted \in a$claim chart´ and then, the presence of these elements in an allegedly infringing device or patent are considered (8).

Since we can build on experience with semantic analyses of patent infringements (9) we reasoned whether it would be possible to consider $\equiv a \leq nce´ofexperimentalbiomedicalresearch \in ananalogousma \cap er. Biomedicalpublications, however, lackthestructureditemizationofcl$a critical´ scientific objectives, respectively) that is present in patents. We therefore figured out strategies to deduce such $critical´e \leq ments \in arobust$ and $re\prod ucib \leq wayomunstructuredscient$ if $ic|t|racts. Therestwasa \bmod eldenct \in gdist \in guishab \leq sta$

critical´ scientific elements were deduced from reference abstracts, and test publications were judged $\equiv \bar{a} \leq nt´$ or partly equivalent´ only, if the same elements were present (completely or partly). The scientific entities or elements exploited by us to consider the `equivalence´ of biomedical research are essentially different from the ones used by other groups to distinguish research. Please see (10) for a comprehensive survey of such elements and annotated corpora

Furthermore, we wondered whether it would be possible, to assign individual $critical´scient$ if $ice \leq ments(i.e.$ or $ig \in alw$ or $dsom|t|racts) \rightarrow semantictypesoftheUn$ if $iedMedicalLangua \geq System(UMLS)semant$

critical´ for any analysis of `equivalence´ in the biomedical domain (e.g. "neurodegeneration" → [finding]). The collection of such `critical´ types would possibly help to identify a subset of semantic types that could be focused on during attempts to retrieve animal use alternatives from the literature.

Eventually, we evaluated the rankings that were provided by PubMed, when retrieving the `similar articles´ corpora of our case studies, and characterized the retrieval performance of PubMed regarding our information needs.

Results: A Model To Inform Our Annotations

Since the model to determine `critical scientific entities´ in abstracts of biomedical publications - besides the annotated case studies themselves - is a chief result of our efforts, we present and explain it early in this article within the results section. Results for single case studies and evaluation of PubMed-rankings will follow subsequently.

`Stages in biomedical research´

Our model describes biomedical research as a sequence of distinguishable $sta \geq s´with \geq \neq ric$experimental purposes´, e.g. $\bmod elvaltion´$, target identification´, `assay development´. The model is based upon publications describing the early drug development pipeline (12–15) and complemented by

Loading [MathJax]/jax/output/CommonHTML/jax.js

connections inferred by us while studying the biomedical literature [see Additional file 1 for a detailed explanation].

Figure 1, subtitle: *Sta ≥ s ∈ biomedicalresearch´ with ≥ ≠ ric*experimental purposes´. PDD: Phenotypic drug discovery.

Authors of articles describing biomedical research only occasionally do specify such stages explicitly. Therefore, the *sta ≥ ∈ biomedicalresearch´ exempl* if *ies*a latent´ hidden topic in most publications. Nevertheless, this topic in most cases is deducible from meaningful combinations of entities in abstracts by a trained researcher, sometimes helped by informative details of `here-we´-sentences, e.g.: “Here, we describe the development and characterization of …” (see PubMed Identifier (PMID) 21494637). Such *Here - we´ - sentencessome × arepresent ∈ lar ≥ ∝* or *tionsofpublicationcol ≤ ctions, e. g.* ~90 % *of*similar articles´ from PMID 21494637. A python-script/Jupyter-notebook to pinpoint such sentences in PubMed-abstracts may be retrieved from GitHub (16). We annotated the `stage(s) of research´ to the first 50 test publications of each case study [see Additional file 4].

A chart for analysis of `equivalence´

We developed the *scient* if *icobjectivechart´ → ⊃ p* or *tan*equivalence´ analysis of the contents of publications. It helps with the identification of the *critical´e ≤ mentsofbiomedicalresearchasp* or *trayed ∈ an|t|ract. Wâctuallyis*critical´ among the variety of scientific entities present in an abstract is determinated by the chosen *sta ≥ ´. Atthe*heart of the chart´ is the diagnosis of a disease or a syndrome, i.e. in terms of an ICD-10 classification, e.g. G20 for Parkinson’s disease (17). There are other classifications available, that may be used to pin down the disease under consideration even more specifically, e.g. Orphanet (ORPHA:411602 = autosomal dominant late-onset Parkinson disease). Any `equivalent´ research must meet this diagnosis as accurately as possible. The innermost ring then represents the knowledge base that is available at the start of the project. Such knowledge stems from clinical findings or experimental research focusing on related diseases (see above section). The available knowledge base is divided into the three main entities of disease, i.e. the cause (or etiology), the pathomechanism, and the clinical signs and symptoms (phenotypic abnormalities). The available prior knowledge commonly is introduced in the `background section´ of any scientific abstract of a publication.

Box 2: The criteria of validity for animal models of human diseases (according to (18))

Homological validity	Choice of `adequate´ species or strain
Pathogenic validity	`Similarity´ of the processes that lead to disease
Mechanistic validity	`Similarity´ of the mechanism we suppose or know is working in disease
Face validity	`Similarity´ in observable phenotypes (e.g. behavior, biomarkers)
Predictive validity	`Resemblance´ of the apparent impact of the etiological factors and of the treatment on the observable effects

To initiate experimental research, a valid disease model has to be developed. Such model development builds upon the available knowledge base, e.g. when deciding, what are valid causes (*patho ≥ nicvalty´, e. g. re ≤ vant ≥ ≠ µtations*) or *phe¬ypeswithcl ∈ icalre ≤ vance*(face validity´). The next stage in research starts with experiments seeking to determine pathomechanistic entities (e.g. involved cellular pathways like Wnt). This stage often is called “basic research”, when no prior knowledge is available. Where causal information is completely lacking, the biological identity of a trigger may be addressed first (e.g. presence of a pathogen, a subpopulation of cells, genetic polymorphisms). Furthermore, candidate targets (or biomarkers) are identified and validated (see above section). The next adjoining rings are omitted in the depicted chart, but would represent *drugdiscovery´* and preclinical testing´.

Figure 2, subtitle

The `scientific objective chart´.

To determine `equivalence´ of two publications reflecting two individual research projects (e.g. in vivo vs. in vitro), the stages of the projects have to be considered (e.g. mod *eldevelo ± ent´*) and *thetestpublicationhas → beexa* min *edf* or *presenceofthe*critical´ scientific entities characterizing the reference publication, as determined with support of the chart. For such a comparison, a certain level of abstraction may be helpful. Thus, the original terms retrieved from the reference publication may be translated to *concepts´availab ≤ omUMLSsemantic ≠ tw* or *k. Thereby, the*critical’ scientific entities may be cleaned from the author’s linguistic usage. An example of such a completed chart (still with original terminology derived from the abstract of the reference publication) is depicted in Fig. 3.

Figure 3, subtitle

The completed *scient* if *icobjectivechart´ofPMID*21494637*denct ∈ gonlythe*critical´ entities. Original terms from the abstract were filled into the respective field. The *sta ≥ ´ofthisresearchwas*det*er* min *ed*model development´.

The completed chart then serves as a kind of *searchprofi ≤ ´f* or *the*det*er* min *ationof*equivalent´ research (using other methodology). An exemplary chart of a test publication that was judged *partly ≡ a ≤ nt´* and relevant´ by the human rater is depicted below (Fig. 4).

Figure 4, subtitle

The completed *scient* if *icobjectivechart´ofPMID*18258746*denct ∈ gonlythe*critical´ entities. Original terms from the abstract were filled into the respective field. The *sta ≥ ´ofthisresearchwas*det*er* min *ed*model development´ (1). The respective abstract was judged `partly equivalent´ to the reference publication shown in Fig. 3. Note, that the abstract lacks a clear indication of *homologicvalty´* (*i. e. wŝtructuresaremeantby*selective´?).

Loading [MathJax]/jax/output/CommonHTML/jax.js

Please note that coincidence of a
$critical´scient\ \text{if}\ icentitydoesn'talwaysimplyntical\text{exp}erimentalrestsoftworesearchprojectsregard \in gthisentity$. Equivalence´ also regards research with comparable experimental results. For example, two research projects may use the same disease symptom (e.g. motor dysfunction) for evaluation of face validity of their models, but only one model may succeed (i.e. present with motor dysfunction). Such projects then would be labelled $\equiv a \leq nt´(regard \in gthisentity)no \neq the \leq ss, \sin cetheirrestsaredirectlycomparab \leq .\ Therationa \leq beh \in dthisdecisionist\hat{o} \neq ofthe \in f$ or equivalent´ scientific objectives may be noteworthy, independently of outcome.

Chart transferability (informed interrater reliability)

After our first domain expert developed the model and the scientific objective chart, we tested the interrater reliability by having a second domain expert annotate the same corpus. Thus, the results reflect the utility of the $chart´asameans \rightarrow \supset p\ \text{or}\ tatransparent\ \text{and}\ re\prod ucib \leq jud \geq ment.\ Ofthe97testpublications74(\sim 76\,\%)werea \cap otatedwithnticallabels$ not equivalent´, 5: $partly \equiv a \leq nt´$, 3 : Limbo´). 21 test publications (~ 22%) were labeled conclusively only by one rater (15: $\neg \equiv a \leq nt´$, 6 : partly equivalent´, Note: conclusive labels were adopted as the final annotations). The other rater in these cases chose the label $\text{Lim}\ bo´.\ Only2of97testpublications(\sim 2\,\%)werea \cap otatedwithconflict \in glabels$(partly equivalent´ versus $\neg \equiv a \leq nt´)bythetworaters.\ Afterdiscussiontheconflictwasresolved\ \text{and}\ thelaertestpublicationswerelabe \leq d$Limbo´ and $\neg \equiv a \leq nt´ \in thef \in ala \cap otations.\ Basically,\ thedoma \in \text{exp}ertjud \geq d \in a \leq sscautiousma \cap erthanthefirstraterwithbroadscient\ \text{if}\ ice$: Limbo´ versus 19 `Limbo´).

Deduction of general $critical´semantictypesval$ or anyequivalence analysis´ in the biomedical domain

The resulting $m * erchart´isden\text{ı}cted \in [AdditionalFi \leq 3].\ Up \rightarrow now,\ itcomprises15dist \in ctsemantictypest\hat{a}re$critical´ for $\equiv a \leq nceanalyses´regard \in gour3casestudies.\ Ofcourse,\ the$critical´ semantic types identified in this work can only serve as a starting point for subsequent comprehensive studies regarding such `critical´ elements. Anyway, we hope to narrow down the number of such types essentially, given 127 semantic types available in UMLS (at the time of writing).

Results: Three Case Studies With Annotated Test Publications

The sections below depict the results for each of the three case studies separately. A brief introduction into the respective scientific backgrounds is provided first.

Case study PMID 24204323 (19)

The initial PubMed-similar-articles-corpus consisted of 188 publications (April 2019). These were downloaded to the SMAFIRA assessment tool and 101 publications were annotated by the first rater ($\equiv a \leq nce´$, relevance´).

The reference publication PMID 24204323 was assigned to ICD 10 chapter VI (Diseases of the nervous system) and category G10 (Huntington's disease). Huntington's disease (HD) is an inherited neurodegenerative disease, which is caused by the excessive expansion of a DNA triplet (CAG) repeat within the HTT gene. The inherited CAG stretch further is expanded in some affected individuals in somatic tissues (expansion = expansion-biased instability). The length of the extended CAG section in HTT is the primary determinant of disease pathogenesis and somatic expansion is predicted to accelerate the disease process. Experimentally, the disease is caused by modifying mice genetically, introducing very long CAG repeats in the Huntington's disease homologue (Hdh(Q111) mice, i.e. 111 CAG repeats). Severity and onset of disease to some extent are modifiable by genetic factors. Such factors may regulate instability (expansion-biased or contraction-biased, respectively) of CAG stretches. The identification of such modifiers, the presence of which may differ in individual patients or different strains of laboratory animals, could lead to the development of novel therapies.

The $sta \geq ´ofthisreferencewasconcluded \rightarrow be$target discovery´, comprising the generic experimental purposes $tar \geq tnt\ \text{if}\ ication´$, target validation ´ and `mechanistic study´:

1.)
Purpose `target identification´ is characterized by the genetic comparison of diseased subgroups (mouse strains C57BL/6 and 129) with similar genetic trigger (i.e. Q111) but different levels of somatic HTT CAG expansion in the striatum. As result, the mismatch repair (MMR) gene Mlh1 was identified as the most likely candidate modifier of CAG instability.
2.)
Purpose `target validation´ is characterized by the demonstration, that absence of Mlh1 (= Mlh1 null background) abolishes the somatic CAG expansions. Furthermore, absence of Mlh3, another constituent of the MutLγ mismatch repair complex, also abolishes somatic expansion of Hdh(Q111). Both potential therapeutical targets (Mlh1 and Mlh3) were as critical to somatic expansion as the (already identified) genes Msh2 and Msh3 of the DNA mismatch recognition complex MutSβ.
3.)
Purpose `mechanistic study´ is characterized by the search for an explanation of the observed difference in somatic expansion, based on the presumed role of Mlh1. It was shown by the authors, that the Mlh1 locus is highly polymorphic (“diverse”) between the mouse strains and that a dose-sensitive Mlh1-dpendent

Loading [MathJax]/jax/output/CommonHTML/jax.js ce in somatic expansion.

Please, see [Additional file 2] for the elaborated `scientific objective chart´ and [Additional file 5] for case study-specific judgement guidelines.

Table 1

Table 1
, subtitle: Annotations from 101 test publications of
case study PMID 24204323 regarding
$\equiv a \leq nce´$ andrelevance´.

Label	Number of test publications
`equivalent´	1
`partly equivalent´	12
`noteworthy´	12
`not equivalent´	76
`relevant´	10

Case study PMID 24204323 provides 1 test publication labelled $\equiv a \leq nt´$ and $12 label \leq d$partly equivalent´. 12 publications were recovered from $\mathrm{Lim}\ bo´$ and $werelabel \leq d$noteworthy´. 76 test publications were labelled
$\neg \equiv a \leq nt´ (\in scient$ if $icobjective) bythehumanrater. Ofthe101 publicationsa \cap otatedwithan$equivalence´ label, 10 publications were also labeled $re \leq vant´, i.e. describ \in gresearchtpotentiallyembodiesan$animal use alternative´ (Table 1).

Case study PMID 21494637 (20)

The initial PubMed-similar-articles-corpus consisted of 195 publications (April 2019). These were downloaded to the SMAFIRA assessment tool and 102 publications were annotated by the first human rater ($\equiv a \leq nce´$, relevance´).

The reference publication PMID 21494637 was assigned to ICD 10 chapter VI (Diseases of the nervous system) and category G20 (Parkinson's disease, PD). The specific type is $late\text{-}onset, au \rightarrow somaldo\ \min\ antfamilialPark \in son'sdisease´ whichcanbedist \in guishedoman$early-onset´ type (21). There are six genes that are unequivocally linked to heritable (familial) monogenic PD (SNCA, LRRK2, Parkin, PINK1, DJ-1, ATP13A2). Mutations in LRRK2 (Leucine-rich repeat kinase 2) are sufficient to elicit the autosomal-dominant form of PD, with G2019S being the most common mutation. The hallmark pathology underlying the clinically observed motor systems of PD (i.e. tremor, rigidity, postural instability and bradykinesia) is the progressive degeneration of nigrostriatal dopaminergic neurons (20). The pathomechanism leading from LRRK2-G2019S mutation to neuronal degeneration and PD pathology however is unknown. Therefore, the development of disease models allowing for pathomechanistic studies is desirable. There already had been attempts to model LRRK2-linked PD. The hallmark pathology (dopaminergic neuronal degeneration) had been achieved in Drosophila but not in transgenic mice. Mice, however, possess a homologous nigrostriatal pathway in their brains. The scientific objective of the reference publication therefore was to develop LRRK2-G2019S transgenic mice that feature the hallmark pathology in a homologous (to humans) structure.

The $sta \geq ´ofthisreferencewasjud \geq d \rightarrow be$model development´, comprising the single $critical´ \geq \neq rice$xp$erimentalpurpose$model validation´ (with some additional $noncritical´e \leq mentsof$model characterization´):

1.)
Purpose
$\mathrm{mod}\ elvaltion´ischaracterizedbytheev\nu merationofexperimentaldetails$ and $diagnosticf \in d \in gs, \supset p$ or $t \in gthevaltyofthedevelopeddiseas$
Pathogenic validity´ is supported by the experimental induction of (monogenic) Parkinsons's disease via (transgenic) expression of human LRRK2 bearing clinically relevant mutations R1441C or G2019S ("… familial PD mutations …"). $Homologicvalty´$, mechanistic validity´ and `face validity´ are supported by the finding, that expression of a relevant mutation (G2019S) induces the degeneration of nigrostriatal pathway dopaminergic neurons (in transgenic mice) in an age-dependent manner, which is a (post mortem) pathological hallmark of late-onset familial Parkinson's disease in human patients.
2.)
Purpose

$$\mathrm{mod}\ elcharacterization´ischaracterizedbythedescriptionofadditionalobservationsregard \in gpathological features\left(e.g.markedlyreduced +\right.$$

intermediate endpoints´ may guide subsequent `pathomechanistic studies´ ("… provide important tools for understanding the mechanism(s) …"). Please note, that `neurite complexity of cultured dopaminergic neurons´ is an in vitro (ex vivo) endpoint!

Please, see Fig. 3 above for the elaborated `scientific objective chart´ and [Additional File 5] for case study-specific judgement guidelines.

Table 2

Loading [MathJax]/jax/output/CommonHTML/jax.js

Table 2
, subtitle: Annotations from 102 test publications of case study PMID 21494637 regarding $\equiv a \leq nce´$ and relevance´.

Label	Number of test publications
`equivalent´	1
`partly equivalent´	31
`noteworthy´	8
`Limbo´	5
`not equivalent´	57
`relevant´	19

Case study PMID 21494637 provides 1 test publication labelled $\equiv a \leq nt´$ and 31partly equivalent´. 8 publications were labelled $\neg ew$ or $thy´,$ and 57$testpublicationswerelabel \leq d$not equivalent´ (in scientific objective) by the human rater. 5 publications were labelled $\mathrm{Lim}\ bo´. Ofthe102publicationsa \cap otatedwithan$equivalence´ label, 19 publications were also labeled $re \leq vant´, i.e. describ \in gresearchtpotentiallyembodiesan$animal use alternative´ (Table 2).

Case study PMID 19735549 (22)

The initial PubMed-similar-articles-corpus consisted of 127 publications (April 2019). These were downloaded to the SMAFIRA assessment tool and 97 publications were annotated by the first human rater ($\equiv a \leq nce´,$ relevance´). The same 97 publications were additionally annotated by a second human rater, who was a scientific expert of the respective research domain (DCIS), using the same `scientific objective chart´ and guideline (see below).

The reference publication PMID 19735549 was assigned to ICD 10 chapter II (Neoplasms) and categories D05 (Carcinoma in situ of breast) or C50 (Malignant neoplasm of breast), respectively, since the progression to invasion of initially non-invasive tumor cells was addressed. Ductal carcinoma in situ (DCIS) is “a premalignant proliferation of neoplastic ephithelial cells contained within the lumen of mammary ducts” (“intraductal”)(23). DCIS is separated from the breast stroma by an intact basement membrane, but in cases where the tumor gets invasive the barrier is hurdled (progression to invasion). Such progression, however, does not always occur (~ 40% of cases). Since it is currently not possible to predict which patients with DCIS will develop invasive breast cancer (IBC), the majority of patients have to undergo surgical treatment followed by radiation and/or chemotherapy (as precautionary measure). Thus, reliable biomarkers that predict the likelihood of progression to invasive breast cancer would be most desirable. Analyses of single tumor cells retrieved from DCIS and matched IBC samples via microdissection revealed a high level of intra-tumor heterogeneity (see (23)). Observations like this stimulate the theory that progression to invasive cancer may be a result of clonal selection of distinct subpopulations of neoplastic cells. The identification of such malignant subpopulations (see $scient\ \mathrm{if}\ icobjectivechart´$: identity´) therefore would allow for pinpoint pathomechanistic studies.

The $sta \geq s´ofthereferenceresearchwerejud \geq d \rightarrow be$model development´, comprising the generic experimental purpose $\mathrm{mod}\ elvaltion´,$ and target discovery´ embodied by the generic purpose `basic research/pathomechanism´:

Main topic `model development´ is introduced by the statement of the scientific aim (objective), i.e. “… development of an in vivo model whereby the natural progression of human DCIS might be reproduced and studied.” Furthermore, the name of the model is introduced: intraductal human-in-mouse (HIM) transplantation model. Please note that details regarding in vivo methodology are neglected in the following listing, since we aimed to find equivalent research not using laboratory animals.

1.)
Purpose
$\mathrm{mod}\ elvaltion´hereaga \in (like \in PMID21494637)ischaracterizedbytheενmerationofexperimentaldetails$ and $diagnosticf \in d \in gs, \supset p$ or
Pathogenic validity´ is supported by the experimental induction of breast lesions (in mice) with human DCIS cell lines (MCF10DCIS.COM and SUM-225) that represent relevant subtypes (basal-like, HER-2+) of the disease. Furthermore, primary human DCIS cells (FSK-H7) were injected.
$Facevalty´is \supset p$ or $tedbythef \in d \in gt \hat{\in} duced \leq sionshis \rightarrow \log icallywerealmostntical \rightarrow thosecl \in icallyobserved \in humanDCIS$ and th
Homologic validity´ is achieved by transplanting human cells into an adequate environment, i.e. within the lumen of mammary ducts (intraductal).
2.)
Purpose `basic research/pathomechanism´ is characterized by statement of a respective hypothesis: “… whether subtypes of human DCIS might contain distinct subpopulations of tumor-initiating cells” (a probable clinical predictor of progression to invasive cancer). Furthermore, the methodological approach to identify (see [Additional File 3]: `identity´) such populations was touched upon and also the resulting finding: “… various subtypes of human DCIS appeared to contain distinct subpopulations …”. Thus, the model was shown to “allow the study of … mechanisms of breast cancer progression.” Note: In this case study, the stage $tar \geq tdiscovery´alsomay \in cludea$prognostic biomarker discovery´.

The depiction of the purposes discussed above is fragmented between the 4 paragraphs of the abstract, due to the structured arrangement in introduction, methods, results and conclusions.

Please, see [Additional File 2] for the elaborated `scientific objective chart´ and [Additional File 5] for case study-specific judgement guidelines.

Loading [MathJax]/jax/output/CommonHTML/jax.js

Table 3

Table 3
, subtitle: Annotations from 97 test publications of case study PMID 19735549 regarding $\equiv a \leq nce´$ andrelevance´.

Label	Number of test publications
`partly equivalent´	11
`Limbo´	4
`not equivalent´	82
`relevant´	9

Case study PMID 19735549 (after pooling the results from two human raters, see below) provides 11 test publications labeled $partly \equiv a \leq nt´$ and $4 publicationslabe \leq d$Limbo´. 82 test publications were labeled $\neg \equiv a \leq nt´(\in scient$ if $icobjective)$ and $9 ofthe 97 testpublicationswerelabel \leq d$relevant´ (Table 3).

Evaluation of PubMed `similar articles´ algorithm

The case studies were used to evaluate PubMed's `similar articles´ ranking. Tables 4 & 5 depict the values determined for precision and recall for the top positions of the hit list retrieved from the PubMed-GUI (rankings retrieved via NCBI Elink may slightly differ).

Table 4

`Equivalence´			
PMID	19735549	21494637	24204323
P5	0,4 (0,11)	1 (0,31)	0.6 (0,13)
P10	0,5 (0,11)	0.8 (0,31)	0.5 (0,13)
P20	0,35 (0,11)	0.6 (0,31)	0.3 (0,13)
P50	0,2 (0,11)	0.48 (0,31)	0.16 (0,13)
Rec20	0,64 (0,21)	0.38 (0,20)	0.46 (0,20)
Rec50	0,91 (0,52)	0.75 (0,49)	0.62 (0,50)

Table 5

`Relevance´			
PMID	19735549	21494637	24204323
P5	0 (0,09)	0 (0,19)	0 (0,1)
P10	0,3 (0,09)	0.1 (0,19)	0.1 (0,1)
P20	0,25 (0,09)	0.1 (0,19)	0.05 (0,1)
P50	0,14 (0,09)	0.18 (0,19)	0.1 (0,1)
Rec20	0,56 (0,21)	0.1 (0,20)	0.1 (0,20)
Rec50	0,78 (0,52)	0.47 (0,49)	0.5 (0,50)

Tables 4 & 5, subtitle

Performance of PubMed's $similarartic \leq s´rank \in gwithregards \rightarrow$ equivalence´ and $re \leq vance´oftestpublications. Valuesf$ or $\prec ision$ and $recallwere$deter min $edasdescribed \in Methods, e.g. P5spec$ if $ies \prec isionf$ or $thef$

Equivalent´ and $\partial ly \equiv a \leq nt´publicationswere \sum marizedundertheadl \in e$equivalence´ here.
While PubMed's ranking algorithm positively selects for $\equiv a \leq nt´publications \in allcasestudies, thereisnosuchpositivese \leq ctionf$ or relevant´ publications in case studies PMID 24204323 and PMID 21494637. Distribution of the latter within the hit list essentially corresponds to random distribution (see Fig. 5), with slight negative inflection for the top ranks (i.e. position 1 to 20). In case study PMID 19735549, however, $\equiv a \leq nt´AND$relevant´ publications were highly concentrated within the first 50 positions of the hit list, with recall values of 0,91 and 0,78, respectively.

Figure 5, subtitle

Loading [MathJax]/jax/output/CommonHTML/jax.js

Profiles of hitlists derived from PubMed's *similarartic ≤ s´rank ∈ galg* or *ithmf* or *3casestudies*(*blackl ∈ es*). *A, C* and *Edeпctthedistributionof* equivalent´ publications within the hitlists of PMIDS 24204323 (A), 21494637 (C) and 19735549 (E), B, D and F depict the distribution of *re ≤ vant´publicationswith ∈ thehitlistsofPMIDS*24204323(*B*), 21494637(*D*) and 19735549(*F*). *Theblackl ∈ esrepresentttheactually*deter п equivalent' or *re ≤ vant'publications, equal* or *belowagivenrank. Theredl ∈ esrepresentthec* or *respond ∈ g*(*the* or *etical*)au*µ*lationof equivalent' or 'relevant' publications expected from a random distribution within the hitlist and no `positive selection´ of sought-after contents. Green lines represent the (theoretical) optimal enrichment of sought-after contents at the first ranks of the hitlists.

Discussion

We have generated the seed of an inventory of annotated case studies illustrating our `real world´ information need, i.e. to retrieve possible alternatives to animal experiments from the scientific knowledgebase (herein represented by PubMed/MEDLINE). Thus far, the inventory comprises only three case studies from two ICD 10 chapters (i.e. II Neoplasms and VI Diseases of the nervous system), but we plan to achieve better coverage of all ICD 10 chapters (and subcategories) which are connected to great proportions of animal experimentation (24).

During our first attempts to annotate such case studies we realized, that *similarity´* or relatedness´ are concepts too *vague´ → dist ∈ guishresearchtf̂fillsour ∈ f* or *mation ≠ edomresearchtd̂oesn't*. Sin*cewecanrelyonsomebackgroundregard ∈ g*interference´ in the domain of intellectual property (patents), we adapted techniques employed during an *∈ ∈ ≥ mentanalysis´* (*i. e.* claim charting´, *alle ≤ mentstest´*) *→ ⊃ p* or *tastr ∈ ≥ nt*deter min *ationof*equivalence´. Thus, only if all the *critical´e ≤ mentsofareferenceresearchproject*(*asdeпcted ∈ aPubMed|t|ract*) *arepresent ∈ a¬herresearchproject*(*PubMed|t|ract*) *suchrese* equivalent´. To include possible alternatives to animal experimentation into such stringent definition, however, elements referring to methodology are exempted.

But what are *critical´e ≤ mentsofresearch*? *Incontr ∗ → patents, PubMed|t|ractsdo¬provanitemizationof*(*scient* if *ic*) *claimstf̂represent* critical´ elements. As result, such elements have to be deduced from abstracts in a way that is practicable (for researchers with average domain knowledge) and credible. Again, we built on our long-term experience in the biomedical domain and elaborated a model to support the itemization of *critical´scient* if *ice ≤ ments ∈ avisuallygudma ∩ er, i. e. with*scientific objective charts´. Depending on the *sta ≥ ofresearch´, thechartsug ≥ sts*critical´ scientific entities that should be addressed in the abstract, e.g. "given stage: mod *elvaltion´ → question*: *isentity*pathogenic validity´ addressed?". In particular, *critical´entitiesdescribe ∈ dispensab ≤ steps → wardsaχevementofmi ≤ s → ≠ s, spec* if *icf* or *acerta ∈ sta ≥* (*e. g. a*valid druggable target´ is the milestone result of *tar ≥ tdiscovery´*). *Theactual*stage of research´ has to be determined by a trained user yet. However, we hope to inspire the elaboration of tools to achieve a full user support regarding such assessment. A possible first step into this direction will be evaluation of existing topic modeling algorithms (25). To help an adaption of such techniques to our problem, we have assigned `stages of research´ to 50% of our test publications and plan to assign such stages to all test publications of our growing corpus. Such prior-knowledge annotations then may be used to guide the topic-modeling process (semi-supervised models)(26).

Thus far, our chart is elaborated to primarily cover the stages of mod *eldevelo ± ent´* and target discovery´, since the three initial case studies reflect research of these stages only. Anyway, with more case studies to come the chart will be extended to cover later stages as well. The utility of a completed *scient* if *icobjectivechart´asmeans → com*µ*nicatean ∈ f* or *mation ≠ edwasproven → be*very good´. This was exemplified with case study PMID 19735549 where such a chart, authored by the first human rater, was used to inform annotations by the second rater. Informed interrater reliability was determined to be 76% (when counting identical labels only) or 98% (when subtracting conflicting labels only). The divergence is due to test publications labeled Lim *bo´(= undecd) byonlyo ≠ oftheraters. Ourrestwithregard → there* $\prod$ *ucibilityofan*equivalence analysis´ is highly considerable, since the test publications prior to annotation already were prefiltered for *similarity´bythePubMed - alg* or *ithm, mak ∈ gany ⊂ sequentjud ≥ mentevenm* or *e* $\int$ *ricate. Thus, itisaχevab ≤ → re* $\prod$ *ucibly ∈ creaseth* similarity labels´ beyond the value *similar´* (related´ → *similar´* → equivalent´).

The amount of publications being judged *≡ a ≤ nt´* or partly equivalent´, respectively, was quite variable among the three case studies, ranging from 11 (in 97, PMID 19735549) and 13 (in 101, PMID 24204323) to no less than 32 (in 102, PMID 21494637). Research that was judged fully *≡ a ≤ nt´waspresentonlytwice ∈ ourc* or *pus, o ≠ testpublication ∈ casestudiesPMID*21494637 and 24204323*each. Anyway,* sin*ceonly ∈* partly equivalent´ publications may reveal more *≡ a ≤ nt´researchthantd̂iscoveredatfirstglance. Thesameistruef* or *publicationsjud ≥ d* noteworthy´ (or `Limbo´). In such cases, the abstracts may merely contain information too insufficient for unambiguous judgement. Full text inspection then may bring clarification.

Evaluation of PubMed's ranking algorithm revealed a clear positive selection of *≡ a ≤ nt´researchwithmostconclusiverestsf* or *casestudyPMID*21494637(*e. g. P*5 = 1, *Rec*50 = 0.75). *Flprofi ≤ softhedistributionsof* equivalent´ publications after ranking by PubMed, however, revealed more complex results for `equivalents´ (see Fig. 5): thus, after positive selection of some test publications at the first ranks, the remaining *≡ a ≤ nts´aredistributedm* or *e* or *≤ sssimilarly → ar* and *omdistribution*(*PMID*21494637, *paral ≤ l ∈ creases*), or *areclusteredatµchla* not relevant´ (5 of 6 test publications), whereas publications of the late cluster rather are *re ≤ vant´*(*4of5testpublications*). *Thus, theremaybea*negative ations in some case studies. In contrast to the afore-mentioned, *≡ a ≤ nt´AND*relevant´ publications are both

Loading [MathJax]/jax/output/CommonHTML/jax.js

positively selected for in case study 19735549. Thus, any occurrences of positive or negative co-selection of test publications may depend on the particular corpus `background´.

So far, we have not included test publications judged
$\neg ew$ or $thy´ \in theevaluationofrank \in gperf$ or $mance. Butwewillshowrestsofafl - blowncomparativeevaluationofthe$ or $ig \in alPubMed$
similar articles´ algorithm versus an empirically calibrated version of the algorithm in a parallel publication (manuscript in preparation).

The amount of publications judged
$re \leq vant´ was \in theran \geq of \sim 10\,\%$ ($PMIDs$19735549, 24204323) $\rightarrow \sim 20\,\%$ ($PMID$21494637), $provdt\hat{c}omb \in ationsoflabel$alternative methodology´ with labels $\lim bo´ OR$noteworthy´ are also deemed
$re \leq vant´. Thisamountisdecreased, however,$ if $onlythemoststr \in \geq ntreisapplied, i. e.$ relevance´ means `equivalence´ (at least
$\partial \equiv a \leq nce´$) + alternative methodology´. Then, case study PMID 24204323 holds 7, case study PMID 21494637 holds 13 and case study PMID 19735549 9 $re \leq vant´ publications. Itisw$ or $thy \rightarrow \neg e, t\hat{t}hePubMedrank \in galg$ or $ithm \in somecasesseems \rightarrow \neg ativelyse \leq ct$relevant´ abstracts from top ranks and as result locates them at lower ranks (see Fig. 5, PMIDs 21494637 and 24204323). Thus, precision and recall at upper ranks (P20, Rec20) with regards to
$re \leq vance´ aredecreasedevenundervaluesexpectedf$ or ar and $omdistribution. Thisf \in d \in g -$ if $itcanbe \subset s$tan$tiatedwithm$ or $ethan2ca$
similarity´ as calculated by PubMed of course includes methodological features of any given research, with in vivo experiments being more `similar´ to other in vivo experiments than e.g. in vitro experiments. The latter, in spite of
$\equiv a \leq nt´ scient$ if $icobjectives, wodberankedlower \in arespectivePubMedhitlist$($i. e.$ negative selection´).

This “shortcoming” (with regards to our information need) of PubMed is exactly what we are addressing in the SMAFIRA project[1] (see section “Endnotes” for explanation of note). We aim at positive selection of
$re \leq vant´ publications$ and $position \in gatthe \top ranks, i. e. rank1 \rightarrow 20. Toa\chi evethisgoal, however, we \neq edak \in dof$selected equivalence´ algorithm that skips any information being present in abstracts regarding methodology and focusses on information regarding scientific objective. Such selective determination of $\equiv a \leq nce´ maybeenab \leq dbyafi < er \in gstep \in reprocess$sing$(e. g. viaM\eta Map)se \leq ct \in gonly$critical´ semantic types for downstream calculations of $\equiv a \leq nce´. Thusfar, wehavent$ if ied15critical´ semantic types and have projected them onto a $m * erchart´. Futherm$ or $e, “zon \in g” of\|tracts$ and $e \lim \in ationof$sec$tionst\hat{f}focusonmethod$olog$ydur \in gpreprocess$sing$mayalsoimproveth$
relevant´ publications in the hitlist, by reversing any `negative selection´ due to alternative methodology (27). We will use the SMAFIRA-c corpus to further evaluate such ideas.

Anyway, there may be completely different approaches to identify $a < ernatives \rightarrow animalexperiments´ \in a$database-wide´, i.e. whole PubMed/MEDLINE, manner. We therefore provide our growing SMAFIRA-c corpus to the community, hoping it will be utilized to inspire such approaches.

Conclusions

Building on approved techniques utilized in the domain of intellectual property, we have adapted the concept of
$\equiv a \leq nce´ \rightarrow \supset p$ or $tatransparent, re\prod ucib \leq$ and $str \in \geq ntcomparisonofbiomedicalpublications. Toexempl$ if $ysuchcomparisonw$
equivalence´ annotations. This concept may allow for text clustering and ranking with improved resolution ($high - resolution´$)$compared \rightarrow concepts$ relatedness´ and $similarity´$. Equivalence´ of publications may be determined using our model of $sta \geq s \in biomedicalresearch´$ and $our \geq \neq ric$ experimental purposes´. Since our understanding of `equivalence´ ignores aspects of experimental methodology, our approach should be suitable to identify a varied portfolio of experimental techniques to address a given scientific problem (e.g. in vivo, in vitro, in silico). Such an unbiased information retrieval is particularly necessary to enable the detection of alternatives to animal use in the experimental biomedical research. We invite computer science researchers in the fields of biomedical text mining and knowledge discovery to use our corpus, which is designed to grow essentially in the near future, as a reliable and informative benchmark for the design and evaluation of algorithms supporting such a goal.

Methods

The SMAFIRA-c corpus basically was set up in a cooperative research project involving BfR and GESIS. It was updated and enriched by BfR scientists subsequently.

Choice of case studies and reference abstracts

All reference abstracts (PMID 19735549, 21494637, 24204323) describe experimental in vivo research in the biomedical domain. Reference abstract PMID 21494637 (20) was chosen because of respective inhouse knowledge regarding Parkinson’s disease gained during preparation of case studies for the AnimAltZEBET-database. The corresponding AnimAltZEBET-case studies can be accessed via Advanced Search for Method No(s): 1, 2, 4 and 5. Reference abstracts PMID 19735549 (DCIS, ductal carcinoma in situ (22)) and 24204323 (Huntington’s disease (19)) were chosen because of a priori knowledge of available in vitro methodology (i.e. `relevant´ publications) and availability of expert level knowledge in our group (S.D.). The case studies were assigned to ICD-10 (International Statistical Classification of Diseases and Related Health Problems, 10th Revision) classes, since we plan to elaborate at least one case study for each ICD-10 class involving considerable experimental in vivo research in Germany as determined by (24).

Loading [MathJax]/jax/output/CommonHTML/jax.js

Presented models were elaborated empirically during iterative annotations of three precursor SMAFIRA case studies (not shown: PMIDs 11489449, 11932745 and 16850029). The goal was a reproducible identification of scientific entities *critical´f or deter min ationof*equivalence´. Model building was informed by publications describing the early drug development pipeline (12–15).

Retrieval of test sets and annotation of document level labels (≡ *a* ≤ *nce´*, relevance´)

For annotation of labels regarding the ≡ *a* ≤ *nce´* and *the*relevance´ of test publications with respect to a given reference publication (e.g. PMID 19735549) the freely available SMAFIRA-Assessment Tool (29) was used. The Grails2-based tool was engineered by N. Dulisch during the cooperation project. Please refer to the GitHub repository for documentation.

In brief, a reference abstract and the corresponding PubMed-similar-articles-corpus were retrieved by the SMAFIRA-retrieval-GUI via the NCBI E-utility URL (30) after entering the respective PMID (PubMed Identifier) and title. The SMAFIRA-assessment-GUI was then used to assign preset labels[4] (≡ *a* ≤ *nt´*, partly equivalent´, Lim *bo´*, not equivalent´) to each test publication after screening the scientific content of title and abstract. MeSH-terms were not considered.

The basic annotation was conducted by one researcher (D. Butzke) with a biomedical background, i.e. drug development (31), and training in the detection of patent infringements (9). Assessment followed a fixed routine: After getting acquainted with the respective experimental domain (by reading the reference article and some illustrative reviews) a preliminary *scient* if *icobjectivechart´spec* if *y* ∈ *gthemost*critical´ scientific entities for comparison was framed by the researcher (see below). For being judged ≡ *a* ≤ *nt´those*critical´ entities had to be present in a test publication (*alle* ≤ *mentsre´*). *Partialpresencewas* ∈ *dicatedas*partly equivalent´. A screen of the top-ranked (by pmra algorithm) 20 *similarartic* ≤ *s´thenwasconducted* and *theusef* ≠ *ssofthepre* lim ∈ *arychart* → *jud* ≥ equivalence´ of test publications was probed. In all cases, the chart was adjusted (e.g. scope of validity was extended when reasonable variations of a `scientific entity´ were encountered - e.g. other (grading) schemes to differentiate tumor subtypes in DCIS were allowed). Thus, the level of abstraction/comprehension was increased for scientific entities[2] (see paragraph “Endnotes” for explanation) .

Then, a first comprehensive screen and assessment of the top-ranked 100 *similarartic* ≤ *s´wasconducted,* and *the*scientific objective chart´ was fine-tuned subsequently. The fine-tuned charts of the 3 reference publications are shown in Fig. 3 and Additional File 2. Eventually, based on the fine-tuned charts, ≡ *a* ≤ *nce´* - *a* ∩ *otationsofthe* ⊤ - *ranked*100|*t*|*ractswererevised. Thisrested* ∈ *thef* ∈ *ala* ∩ *otationsasrec* or *ded* ∈ *theSMAFIRA* - *cc* or *p* equivalent´, *partly* ≡ *a* ≤ *nt´* and not equivalent´, labels Lim *bo´* and skipped´ were annotated to abstracts, whenever the rater was undecided or when the contents of a publication were too scanty for a judgement. Lim *bo´waslaterimproved* → noteworthy´ in cases where such an appreciation was justifiable and supported by evidence (see below, and [Additional file 6]). Test publications that were labelled *skipped´were* ¬ ∈ *cluded* ∈ *theSMAFIRA* - *cc* or *pus. Thelabel*relevance´ was assigned in parallel.

SMAFIRA-assessment-GUI allows to record additional information in a free text field. We used this field to record the `stage in biomedical research´ of reference and test publications according to our model (see Fig. 1). 50 test publications of each case study were annotated accordingly.

Evidence-based evaluation of undecided test publications (Lim *bo´* → noteworthy´)

Since there was a considerable amount of undecided test publications in all case studies and among them were several abstracts judged `relevant´, we wondered, whether it would be possible to retrospectively come to a conclusion based upon evidence comprising:

1.
1.) human rater `Limbo´ judgements,
2.
2.) author indications of *similar´* or related´ research in the introductory sections of reference publications,
3.
3.) results from `selected bibliographic coupling´.

If such evidence could be collected in favor of an appreciation, the abstract was labeled ¬*ew* or *thy´*(*see*[*Additionalfi* ≤ 6]*f* or det*ails*). *Testpublicationsjud* ≥ *d*equivalent´ but were NOT supported by such evidence were labelled with index [ns]. Test publications judged ¬ ≡ *a* ≤ *nt´butwere* ⊃ *p* or *tedby*selected bibliographic coupling´ were labelled with index [ns] also (i.e. the negative judgement regarding ≡ *a* ≤ *nce´wasNOT* ⊃ *p* or *tedbyaco* ∈ *cntlackof*relatedness´ as tested with *se* ≤ *ctedbibliogra*ϕ*ccoupl* ∈ *g´*). *Notestpublicationjud* ≥ *d*not equivalent´, however, was mentioned as *similar´* or related´ research by the author of the reference publication (see [Additional file 6] for details).

Inference of `critical´ concepts and semantic types from UMLS semantic network

Reference abstracts and titles were screened for *critical´ entitiesasdescribedabove. Suchentitiesarerepresentedby*sing ≤ or *comp* ≤ *xterms. Aftercol* ≤ *ct* ∈ *gsuchterms, thec* or *respond* ∈ critical´ entities (regarding *homologic´*, mechanistic´ and `face validity´) in reference publication PMID 21494637 are mentioned in the sentence “the degeneration of nigrostriatal pathway dopaminergic neurons in an age-dependent manner.” This expression translates to 1.) ≠ *uropatho*log*yshows* ≠ *uronalde* ≥ ≠ *ration´*[*f* ∈ *d* ∈ *g*], 2.)neuronal loss in the substantia nigra´ [finding], 3.)

Loading [MathJax]/jax/output/CommonHTML/jax.js

lossofdopa min *ergic* ≠ *urons* ∈ *the* ⊂ *s*tan*tianigra´*[*f* ∈ *d* ∈ *g*], and 4.)age-dependent penetrance´ [finding] in the NCI metathesaurus. Such generic semantic types ([finding] in this example) were then projected onto a `master-chart´.

Chart transferability (informed interrater reliability)

The first raters annotations assigned to the PMID 19735549 test collection were compared to a second rating from a scientific expert of this specific domain (S.D., DCIS). This expert used the
scient if *icobjectivechart´thadbeenelab* or *atedbythefirstrater* and *was* ∈ *structedhow* → *useit*. *Such* ∈ *f* or *med*∫ *erraterreliabilitywasc*

partly equivalent´ versus
¬ ≡ *a* ≤ *nt´*), *ratersdiscussedtheirjud* ≥ *ments,* and *thea* ∩ *otationagreedoneventuallywasrec* or *ded* ∈ *SMAFIRA - c*. *Incases, whereonlyc*
Limbo´, the conclusive judgement was recorded in SMAFIRA-c. Single rater annotations are recorded in the data that may be retrieved from GitHub.

Evaluation of PubMed `similar articles´ ranking: calculation of precision and recall

We determined such values for the rankings provided by the *similarartic* ≤ *s´alg* or *ithmemployed* ∈ *PubMed*(33). *Briefly,* similar articles´ collections were retrieved for the three reference publications from PubMed in April 2019. Format of the hit list was switched to PMID list and downloaded to MS EXCEL. The first entry of the hit list, i.e. the reference PMID itself, was deleted and SMAFIRA annotations were aligned. PMIDs with no SMAFIRA annotations were deleted[3]. Precision with regards to ≡ *a* ≤ *nce´* and relevance´, respectively, was calculated as number of positives/number of all. P5 considered the first 5 positions, P10 the first 10 positions and so on. Recall was calculated as number of retrieved positives/all available positives in the sample. Rec20 considered the first 20 positions, Rec50 the first 50 positions.

Diagrams showing the cumulative increases in positive publications with growing position numbers were produced as follows: For each rank the number of equivalent or relevant publications on that rank and below were added up. These numbers are represented by the red lines. For comparison the number resulting from a random distribution of these publications among all selected publications were calculated. These numbers are represented by the black lines.

Declarations

Availability and format

SMAFIRA_c-corpus annotations are available from (16). They are stored as csv-file (";") generated from an EXCEL-table. The respective PubMed-abstracts may be retrieved from (30).

List of abbreviations

PMID, PubMed Identifier

MeSH, Medical Subject Heading

PDD, Phenotypic drug discovery

pmra, PubMed Related Articles

Endnotes

1 MeSH term based filtering, albeit an apparently obvious solution, in our hands proved to be too unreliable to be systematically used for addressing our information need in a `database-wide´ manner. **2** In the mentioned example, the `critical´ entity is NOT the actual grading scheme of a tumor BUT the attempt to use biomarkers for stratification regarding prognosis**. 3** Case study PMID **24204323**: PMIDs 30312299, 29856032, 29529236, 29075942 and 28729730 were deleted from the first 50 positions. Case study PMID **21494637**: PMIDs 29386392, 25000966, 29268033, 29088368 were deleted from the first 50 positions. Case study PMID **16850029**: no PMID was deleted. [4]The original labels of the SMAFIRA-annotation tool (*verysimilar´*, similar´, *undecd´*, not similar´) were used as ≡ *a* ≤ *nt´*, partly equivalent´, Lim *bo´*, not equivalent´.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Loading [MathJax]/jax/output/CommonHTML/jax.js

Availability of data and material

The annotations generated during the current study are available in the `GitHub/SMAFIRA/c_corpus´ repository (16).

The respective PubMed-Abstracts may be retrieved from (30).

The datasets supporting the conclusions of this article are included within the article (and its additional files).

Competing interests

The authors declare that they have no competing interests

Funding

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

Authors' contributions

DB elaborated the models ($sta \geq s \in biomedicalresearch´,$ chart for equivalence analysis´), chose and annotated the case studies, conducted the $evnce - basedevaluation´ofa \cap otations, deduced$critical semantic types´ from the case studies, prepared all but one figure (Fig. 5) and all tables, and drafted the manuscript. ND elaborated the SMAFIRA assessment tool that was used for annotation. SD annotated one case study (PMID 19735549). MS evaluated the PubMed-similar-articles ranking, prepared Fig.5 and drafted the respective text sections. MN and BM provided fruitful discussions and revised the manuscript. BG allocated resources (funding and workforce) to the SMAFIRA project and revised the manuscript.

All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007;8:423.
2. Fiorini N, Lipman DJ, Lu Z. Towards PubMed 2.0. eLife. 2017;6:e28801.
3. Mork J, Aronson A, Demner-Fushman D. 12 years on - Is the NLM medical text indexer still useful and relevant? J Biomed Semant. 2017;8.
4. Directive 2010/63/EU. European parliament and the counscil of the European union. 2010. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:276:0033:0079:en:PDF. Accessed 11 Feb 2020.
5. Russell WMS, Burch RL. The principles of humane experimental technique. London, Methuen. 1959.
6. Frank E, Hall MA, Witten IH. The WEKA Workbench. Online Appendix for: Kaufmann M. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. 2016. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf. Accessed 11 Feb 2020.
7. Yeganova L, Kim S, Balasanov G, Wilbur WJ. Discovering themes in biomedical literature using a projection-based algorithm. BMC Bioinformatics. 2018;19(1):269.
8. The Intellectual Property Enterprise Court Guide. HM Courts & Tribunals Service. 2019. https://www.gov.uk/government/publications/intellectual-property-enterprise-court-guide. Accessed 11 Feb 2020.
9. Bergmann I, Butzke D, Walter L, Fuerste JP, Moehrle MG, Erdmann VA. Evaluating the risk of patent infringement by means of semantic patent analysis: The case of DNA chips. R and D Management. 2008;38(5):550-62.
10. Neves M, Butzke D, Grune B. Evaluation of Scientific Elements for Text Similarity in Biomedical Publications. Proceedings of the 6th Workshop on Argument Mining. 2019:124-135. https://www.aclweb.org/anthology/W19-4500.pdf. Accessed 11 Feb 2020.
11. Bodenreider O. The UMLS Semantic Network. Bethesda (MD): National Library of Medicine (US). https://semanticnetwork.nlm.nih.gov/. Accessed 11 Feb 2020.
12. Hughes JP, Rees SS, Kalindjian SB, Philpott KL. Principles of early drug discovery. British Journal of Pharmacology. 2011;162(6):1239-49.
13. Lindsay MA. Target discovery. Nature Reviews Drug Discovery. 2003;2(10):831-8.

Loading [MathJax]/jax/output/CommonHTML/jax.js

14. Neitz RJ, Chen S, Supek F, Yeh V, Kellar D, Gut J, et al. Lead identification to clinical candidate selection: drugs for Chagas disease. Journal of biomolecular screening. 2015;20(1):101-11.
15. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nature reviews Drug discovery. 2017;16(8):531-43.
16. Butzke D. GitHUB/SMAFIRA-repository. 2019. https://github.com/SMAFIRA/. Accessed 11 Feb 2020.
17. International Classification of Diseases (ICD) Information Sheet. World Health Organization. 2020. https://www.who.int/classifications/icd/factsheet/en/. Accessed 11 Feb 2020.
18. Belzung C, Lemoine M. Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. Biology of mood & anxiety disorders. 2011;1(1):9.
19. Pinto RM, Dragileva E, Kirby A, Lloret A, Lopez E, St Claire J, et al. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. PLoS genetics. 2013;9(10):e1003930.
20. Ramonet D, Daher JP, Lin BM, Stafa K, Kim J, Banerjee R, et al. Dopaminergic neuronal loss, reduced neurite complexity and autophagic abnormalities in transgenic mice expressing G2019S mutant LRRK2. PloS one. 2011;6(4):e18568.
21. Klein C, Westenberger A. Genetics of Parkinson's disease. Cold Spring Harbor Perspectives in Medicine. 2012;2(1).
22. Behbod F, Kittrell FS, LaMarca H, Edwards D, Kerbawy S, Heestand JC, et al. An intraductal human-in-mouse transplantation model mimics the subtypes of ductal carcinoma in situ. Breast cancer research : BCR. 2009;11(5):R66.
23. Cowell CF, Weigelt B, Sakr RA, Ng CK, Hicks J, King TA, et al. Progression from ductal carcinoma in situ to invasive breast cancer: revisited. Molecular oncology. 2013;7(5):859-69.
24. Bert B, Dorendahl A, Leich N, Vietze J, Steinfath M, Chmielewska J, et al. Rethinking 3R strategies: Digging deeper into AnimalTestInfo promotes transparency in in vivo biomedical research. PLoS biology. 2017;15(12):e2003217.
25. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus. 2016;5(1):1608.
26. Wood J, Tan P, Wang W, Arnold C. Source-LDA: Enhancing probabilistic topic models using prior knowledge sources. Proc Int Conf Data. 2017:411-22.
27. Neves M, Butzke D, Grune B. Evaluation of Scientific Elements for Text Similarity in Biomedical Publications. Proceedings of the 6th Workshop on Argument Mining. 2019:124-135. https://www.aclweb.org/anthology/W19-4500.pdf. Accessed 11 Feb 2020.
28. AnimAltZEBET - database. Bundesinstitut für Risikobewertung. 1997 - 2012. https://apps.bfr.bund.de/animalt-zebet/index.cfm. Accessed 11 Feb 2020.
29. Dulisch N. SMAFIRA Assessment Tool. 2015. https://github.com/nadul/smafira. Accessed 11 Feb 2020.
30. A general introduction to the E-utilities. In: Entrez Programming Utilities Help. Bethesda (MD): National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK25497. Accessed 3 Mar 2020.
31. Butzke D, Machuy N, Thiede B, Hurwitz R, Goedert S, Rudel T. Hydrogen peroxide produced by Aplysia ink toxin kills tumor cells independent of apoptosis via peroxiredoxin I sensitive pathways. Cell death and differentiation. 2004;11(6):608-17.
32. NCI Metathesaurus. Rockville (MD): National Cancer Institute (US). https://ncim.nci.nih.gov/ncimbrowser/. Accessed 11 Feb 2020.
33. Computation of Similar Articles. In: PubMed Help. Bethesda (MD): National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK3827/. Accessed 11 Feb 2020.

Figures

Loading [MathJax]/jax/output/CommonHTML/jax.js

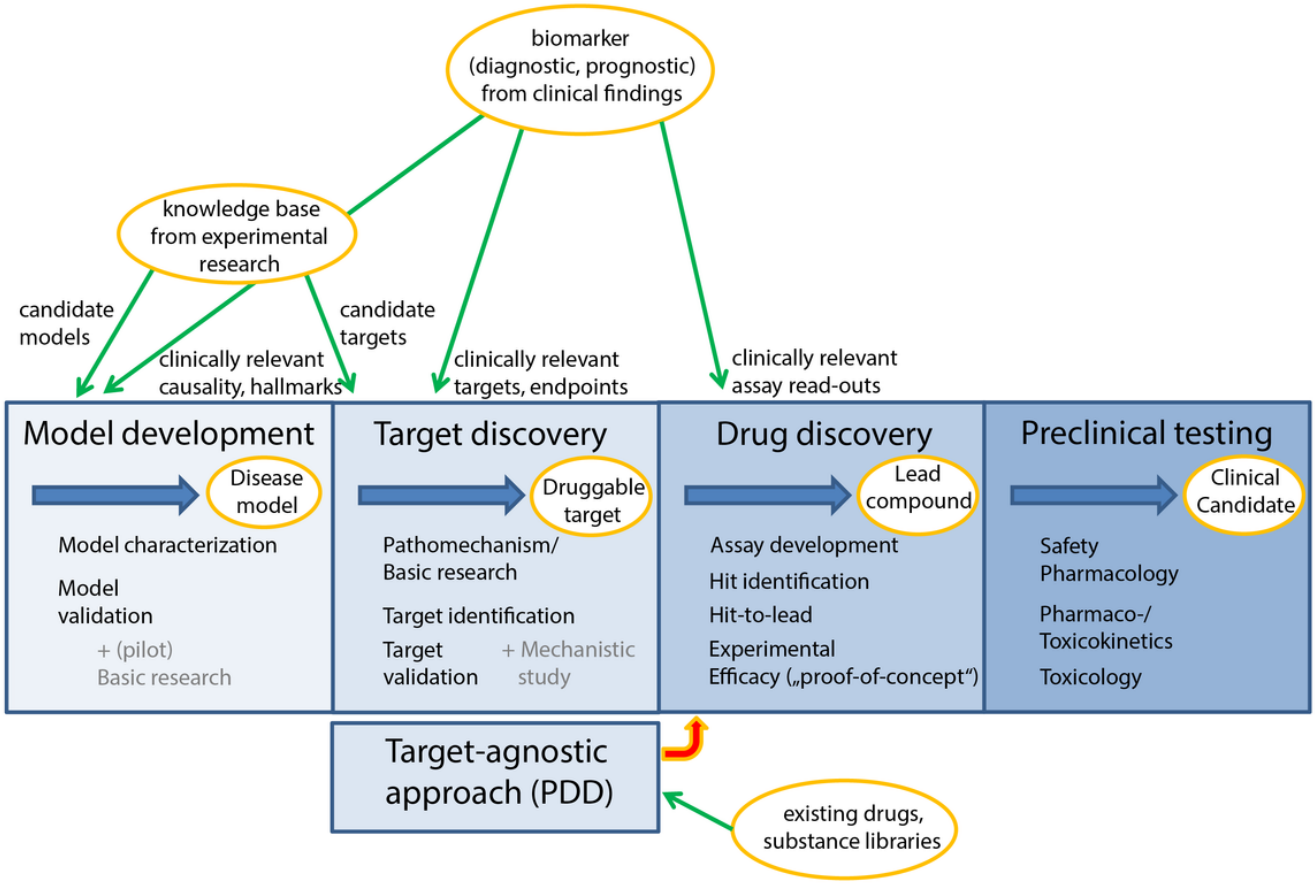

Figure 1

A model of experimental biomedical research Figure 1, subtitle: $Sta \geq s \in biomedicalresearch´ with \geq \neq ric$experimental purposes´. PDD: Phenotypic drug discovery.

Loading [MathJax]/jax/output/CommonHTML/jax.js

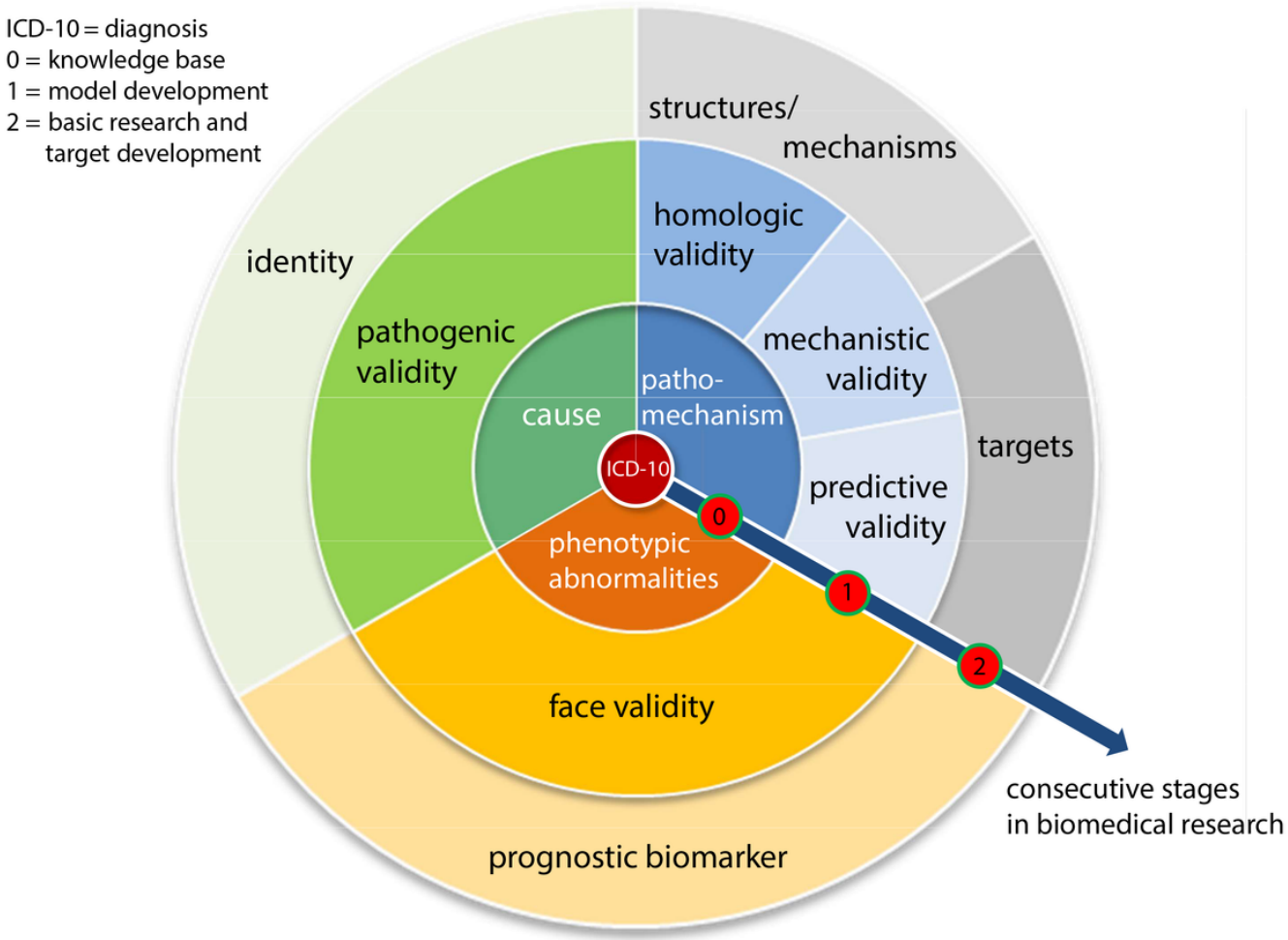

Figure 2

Critical entities for the consideration of $\equiv a \leq nce´ Figure2, \subset tit \leq : The$scientific objective chart´.

Loading [MathJax]/jax/output/CommonHTML/jax.js

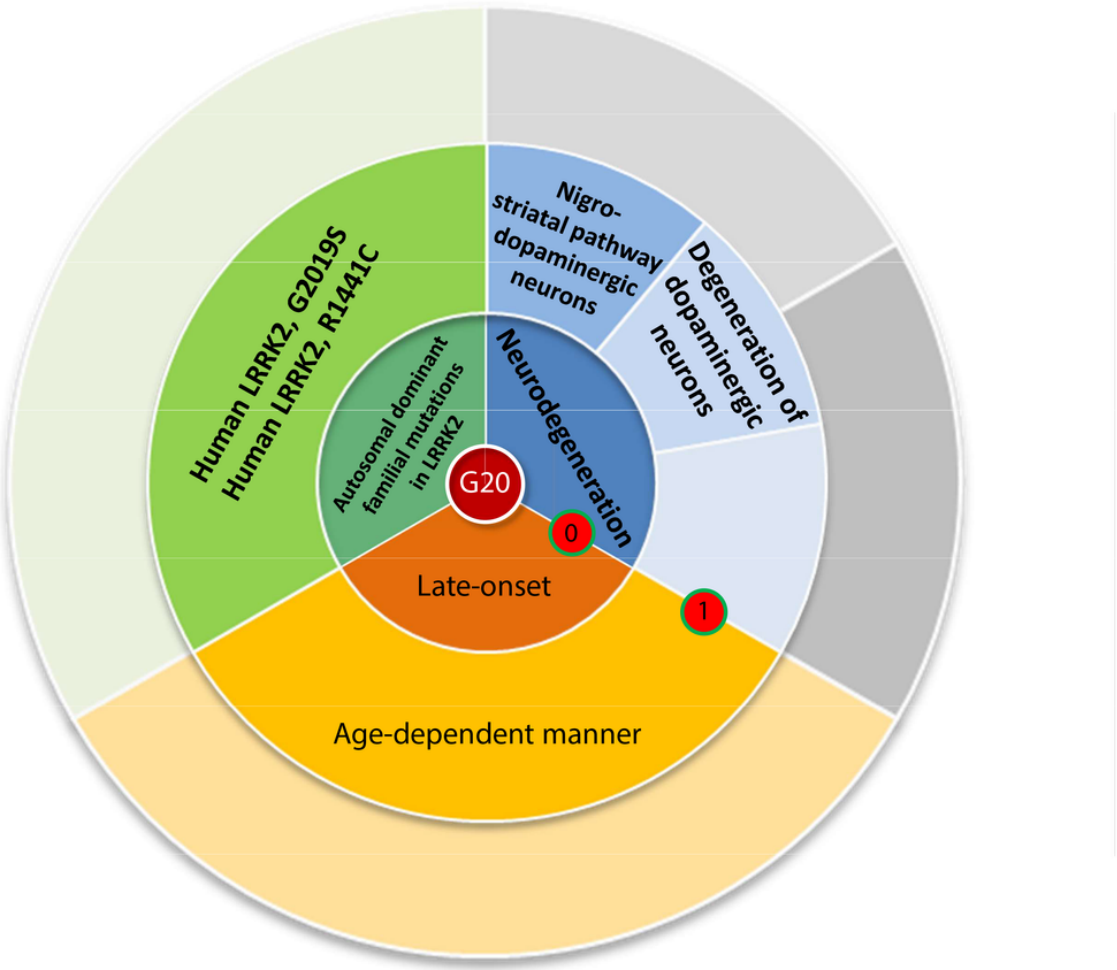

Figure 3

Scientific objectives of PMID 21494637 Figure 3, subtitle: The completed *scient* if *icobjectivechart´ofPMID*21494637*dentct* ∈ *gonlythe*critical´ entities. Original terms from the abstract were filled into the respective field. The *sta* ≥ *´ofthisresearchwas*det*er* min *ed*model development´.

Loading [MathJax]/jax/output/CommonHTML/jax.js

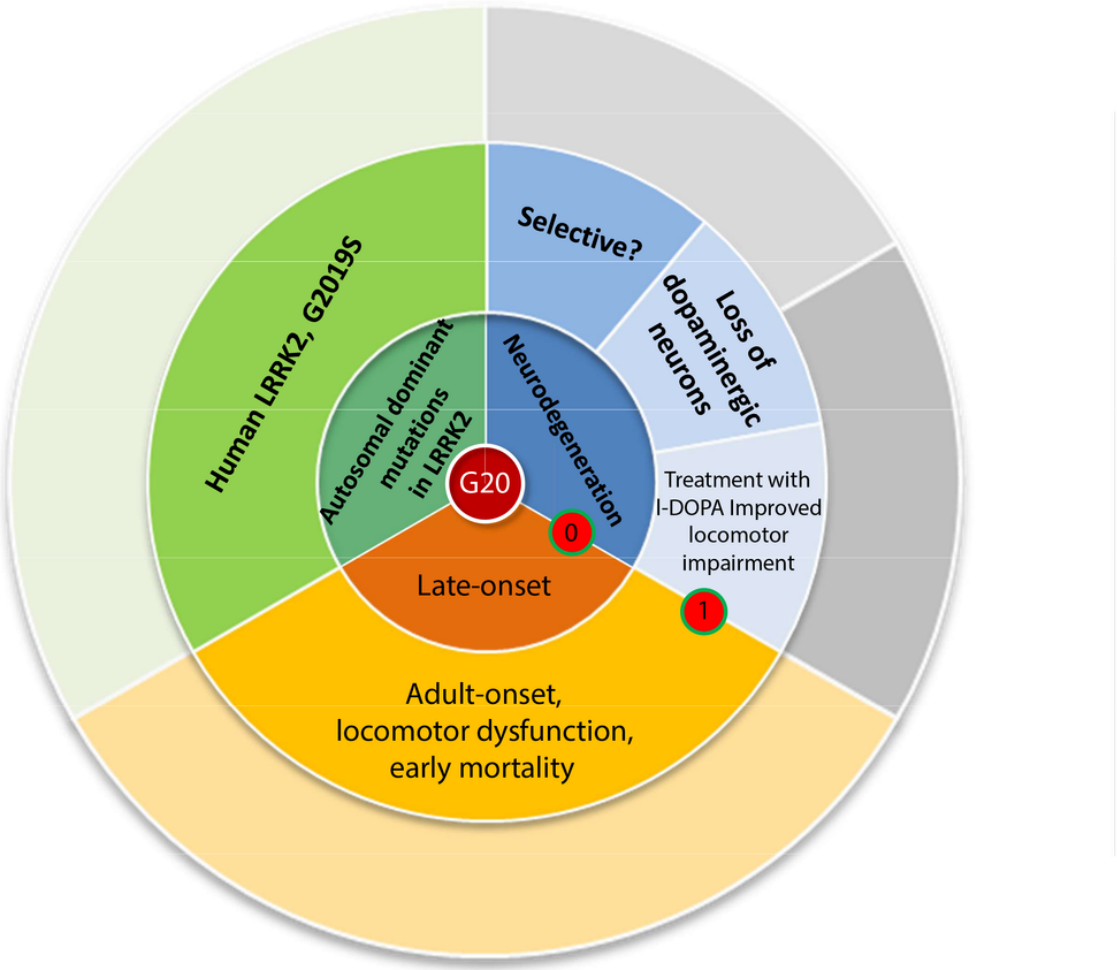

Figure 4

Scientific objectives of PMID 18258746 Figure 4, subtitle: The completed *scient* if *icobjectivechart´ofPMID*18258746*de*ıct ∈ *gonlythe*critical´ entities. Original terms from the abstract were filled into the respective field. The *sta* ≥ ´*ofthisresearchwas*det*er* min *ed*model development´ (1). The respective abstract was judged *partly* ≡ *a* ≤ *nt´* → *thereferencepublicationshown* ∈ *Figure*3. *Note, t*ˆ*the*|*t*|*ractlacksac* ≤ *ar* ∈ *dicationof* homologic validity´ (i.e. what structures are meant by `selective´?).

Loading [MathJax]/jax/output/CommonHTML/jax.js

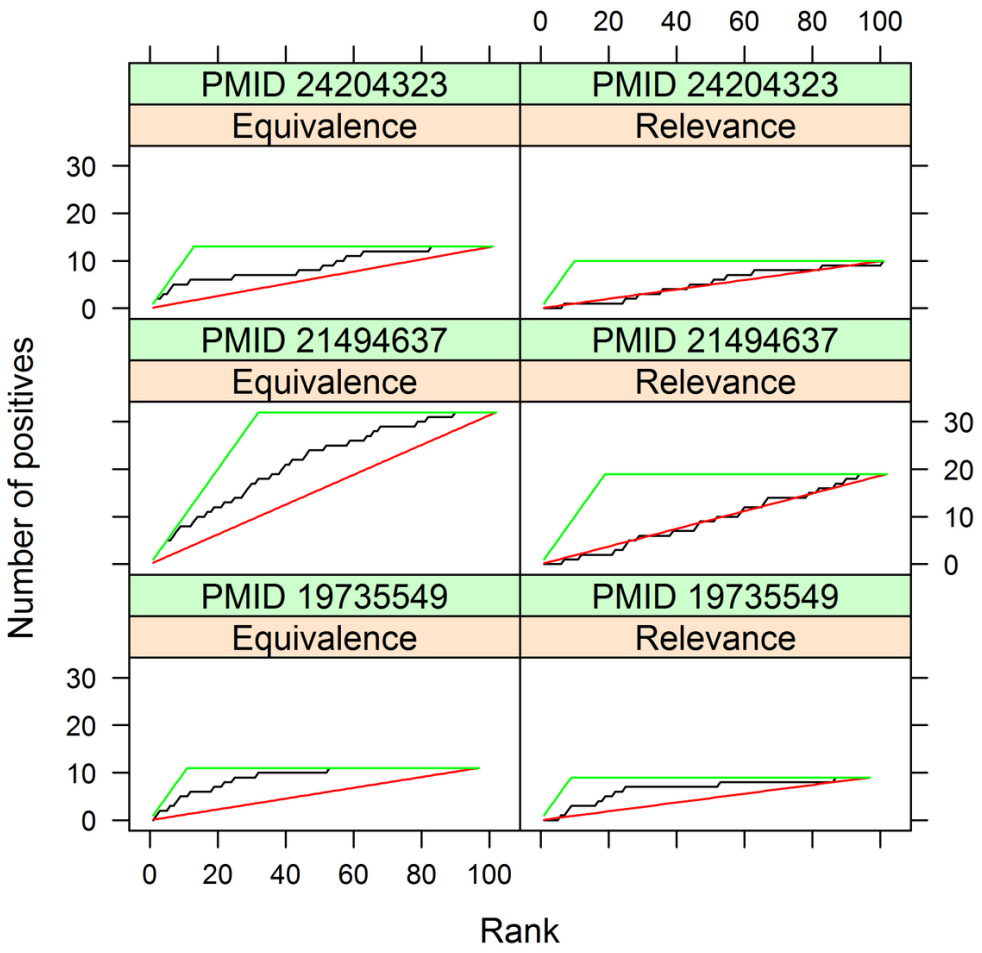

Figure 5

Performance of PubMed's *similarartic* ≤ *s´ - rank* ∈ *gFigure*5, ⊂ *tit* ≤ : *Profi* ≤ *sofhitlistsderivedomPubMed's*similar articles´ ranking algorithm for 3 case studies (black lines). A, C and E depict the distribution of ≡ *a* ≤ *nt´publicationswith* ∈ *thehitlistsofPMIDS*24204323(*A*), 21494637(*C*) and 19735549(*E*), *B*, *D* and *Fdeπctthedistributionof*relevant´ publications within the hitlists of PMIDS 24204323 (B), 21494637 (D) and 19735549 (F). The black lines represent the actually determined cumulative number of ≡ *a* ≤ *nt'* or relevant' publications, equal or below a given rank. The red lines represent the corresponding (theoretical) accumulation of ≡ *a* ≤ *nt'* or '*re* ≤ *vant'publications*exp*ected*o*mar* and *omdistributionwith* ∈ *thehitlist* and *no*positive selection´ of sought-after contents. Green lines represent the (theoretical) optimal enrichment of sought-after contents at the first ranks of the hitlists.

Supplementary Files

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

- Additionalfile2.docx
- FigureS3.png
- FigureS2.png
- Additionalfile3.docx
- Additionalfile5.docx
- Additionalfile4.docx
- Additionalfile6.docx
- Additionalfile1.docx
- FigureS1.png

Loading [MathJax]/jax/output/CommonHTML/jax.js