BMC Bioinformatics

RESEARCH Open Access

Mining impactful discoveries from the biomedical literature

Erwan Moreau[1*], Orla Hardiman[2], Mark Heverin[2] and Declan O'Sullivan[1]

*Correspondence:
moreaue@tcd.ie

[1] Adapt Centre and School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[2] School of Medicine, Trinity College Dublin, Dublin, Ireland

Abstract

Background: Literature-based discovery (LBD) aims to help researchers to identify relations between concepts which are worthy of further investigation by text-mining the biomedical literature. While the LBD literature is rich and the field is considered mature, standard practice in the evaluation of LBD methods is methodologically poor and has not progressed on par with the domain. The lack of properly designed and decent-sized benchmark dataset hinders the progress of the field and its development into applications usable by biomedical experts.

Results: This work presents a method for mining past discoveries from the biomedical literature. It leverages the impact made by a discovery, using descriptive statistics to detect surges in the prevalence of a relation across time. The validity of the method is tested against a baseline representing the state-of-the-art "time-sliced" method.

Conclusions: This method allows the collection of a large amount of time-stamped discoveries. These can be used for LBD evaluation, alleviating the long-standing issue of inadequate evaluation. It might also pave the way for more fine-grained LBD methods, which could exploit the diversity of these past discoveries to train supervised models. Finally the dataset (or some future version of it inspired by our method) could be used as a methodological tool for systematic reviews. We provide an online exploration tool in this perspective, available at https://brainmend.adaptcentre.ie/.

Keywords: Literature-based discovery, Evaluation, Benchmark dataset, Time-sliced method

Introduction

Research papers have been fully digitized for the past 30 years, at least. Paradoxically, while there is no major obstacle to the availability of research, it has never been harder to meaningfully explore the literature due to the sheer amount of publications. Literature-Based Discovery (LBD) aims to automatically extract new insights from the scientific literature [1]. LBD could be of valuable assistance to researchers: by identifying potentially relevant relations between concepts,[1] thus accelerating and broadening scientific progress.

[1] We use the word *relation* throughout this work for any (potential) relationship between two concepts. In general the nature of the relationship is unknown; in theory the association between the concepts does not even have to be positive.

The task of LBD was introduced by Swanson [2]. In this paper, LBD is used to establish a link between dietary fish oil and Raynaud's syndrome, through their known relation to blood circulation. This initial discovery was soon followed by a second one, connecting migraine and magnesium [3]. These two inital discoveries would later become the most commonly used for the purpose of evaluating LBD [4] (occasionally with a few additional discoveries, e.g. in [5, 6]).

Contrary to good practices, LBD evaluation has been based for the most part on this tiny set of discoveries for the past three decades. There are multiple methodological biases in this approach. Double blind experiments are the standard in clinical trials in order to avoid an outcome tainted by the influence of the researcher or the patient. For similar reasons, a reliable LBD evaluation method requires a larger and more diverse sample. Conditions of statistical representativity are not satisfied in the current setting, therefore the results obtained in this way cannot be generalized.

The subpar LBD evaluation methodology might contribute to the lack of uptake by the biomedical research community at large. Despite a rich state-of-the-art, LBD is still a mostly theoretical field. The lack of solid evaluation methodology is probably a factor which hinders the dissemination of LBD as a general research tool.

This is why we introduce the task of mining discoveries from the full existing literature. This task is very similar to LBD in the sense that both aim to produce relevant discoveries as output, but as opposed to LBD it does not have to predict *future* discoveries, i.e. it has access to the data after the time of discovery. By definition this makes the task easier since the system has access to more information. Yet the task is not trivial, because it involves filtering out a lot of relations which do not qualify as discoveries.

In order to formalize the concept of discovery we opt to focus on *impactful* discoveries, i.e. to use the impact of a relation in the literature as a marker of its discovery status: a significant discovery is expected to be followed by a surge in the number of mentions of the relation. We propose a method which calculates the trend across time for a relation based on its frequency in the literature. Significant surges are extracted, leading to a collection of impactful discoveries together with their time of impact. The results of the method are thoroughly analyzed and evaluated against a baseline representing the state-of-the-art "time-sliced" method. The resulting dataset is made available in two forms: the raw data can be used to evaluate or train LBD systems, and the visualization interface facilitates the manual exploration of the data.

The paper is organized as follows: the motivations and main idea are presented in "Approach" section. In "Method" section the method is described in detail, then the experimental results are analyzed in "Results and analysis" section.

Approach

Motivations

Swanson [2, 7] introduced LBD as a method to explore "undiscovered public knowledge", more precisely to identify missing links in a large and fragmented collection of knowledge. His approach, the ABC model, relies on the idea that different fields of specialization tend not to interact with each other. As a result, two subsets of the literature might each contain some knowledge about a shared concept, yet the lack of communication between the two fields can sometimes prevent potentially useful discoveries. In a broader sense, LBD can be defined as a task aimed at generating new research hypotheses, i.e. potential new discoveries [8].

The field of LBD has seen significant progress since its inception. Kastrin and Hristovski [9] systematically analyzed 35 years of LBD literature and observed that the field has grown in volume and diversity, developing from the initial text-based cooccurrences methods to advanced neural-based approaches [6]. However the evaluation of LBD is still a major obstacle to its development into a mainstream research methodology. As previous authors noted, e.g. [4, 10, 11], evaluating LBD is hard due to the nature of the task: there is no obvious way to determine whether the discoveries predicted by a LBD system will eventually turn out to be actual discoveries.

Yetisgen-Yildiz and Pratt [12], as well as Thilakaratne et al. [4] more recently, review the different evaluation methods found in the LBD literature. The *replication* method is still the unrivalled standard in the field: take a well-known discovery at time t, and feed the LBD system only with the literature available before time t (i.e. the established relations at t); then the LBD system is applied, predicting an ordered set of discoveries susceptible to happen after time t. Among these predictions, the likelihood of the initial well-known discovery is an indication of how well the LBD system performs. Naturally, this process should be repeated for multiple discoveries in order to obtain a statistically reliable measure of performance. Nevertheless, it is very common in the literature to evaluate LBD systems against a small number of discoveries. Moreover, the same set of discoveries is used again and again, in particular those by existing LBD methods (typically the ones proposed by Swanson [2, 3]). As mentioned early in the evolution of the field by Ganiz et al. [10], there are several biases in this evaluation methodology:

- There is a risk of confirmation bias, since the method is evaluated in terms of how well it retrieves discoveries that LBD methods are known to be good at finding.
- There is a risk of data leakage (including direct or indirect information from the test set into the model) when a new LBD system A is designed to improve over an older system B and both are evaluated on dataset D, especially if D consists of only a few instances. This can lead to overestimating the performance of system A.
- The performance obtained by evaluating on a small test set lacks statistical reliability, and there is no way to measure the variance in performance. In other words, adding or removing an instance from the test set might affect the overall performance. This makes any comparison between LBD methods fragile.

In [13], a new method was introduced to evaluate a LBD system. This method was later represented as a more formal evaluation approach by Yetisgen-Yildiz and Pratt [11], called *time-sliced evaluation* in [4]. The main idea is similar: define a cut-off year t, and take a term x as the main target; terms which cooccur with x after time t (but not before t) are considered as gold standard discoveries. In other words, the LBD system is expected to find as many relations (x, y) as possible, where y only cooccurs with x later than time t. This solves several issues of the replication method: no cherry-picking discoveries, large sample size, and it also handles negative instances, allowing the use of standard quantitative evaluation mesures such as precision, recall and F-score. This method was adopted for example by Lever et al. [14].

Although originally introduced with discoveries simply defined as the set of cooccurrences found in the literature, the time-sliced evaluation can potentially be used with any source of discoveries, for instance any biomedical database of relations (provided it is known which relations were known before and after the cut-off year). However so far LBD research has only used the time-sliced method with cooccurrences, to the authors' knowledge. In the remainder of this paper, we use the term *time-sliced evaluation* to mean *cooccurrence-based time-sliced evaluation with a large number of instances* for the sake of conciseness.

The time-sliced method solves the serious issue of the small sample size, since most targets have many cooccurrences. However, the asssumption that the set of cooccurrences is equivalent to the set of discoveries is an obvious simplification: actually, a very small proportion of cooccurrences represent actual discoveries. Two concepts frequently appear together only by chance. Additionally, some cooccurrences involve relations between concepts which do not qualify as discovery, such as hypernymy (e.g. *Neurodegenerative Disease* and *Amyotrophic Lateral Sclerosis*) or trivial associations. Spurious relations can be filtered out using a frequency threshold, because very frequent concepts are less likely to be involved in a true discovery. But this coarse treatment is imperfect, and anyway there is no simple solution for the case of chance cooccurrences. Thus there is no clear condition to assess whether a relation should be considered as discovery or not.

To sum up, the time-sliced evaluation is based on a large but very noisy sample. This sample does contain true discoveries, but they probably represent only a drop among a sea of cooccurrences. Therefore interpreting random cooccurrences as gold standard cannot lead to a reliable performance mesure. More precisely, such an evaluation reliably measures the ability of a system to predict cooccurrences, but not to predict insightful discoveries. It follows that, even though the time-sliced evaluation method fixes several serious issues with the replication method, it is still not fit for purpose.

LBD is by nature data-driven, and the standard methodology is to evaluate data-driven tasks against benchmark datasets. Usually a benchmark dataset is adopted by a community if it is seen as representing the task (or some aspect of it) faithfully. This allows systems for this task to be evaluated and compared on the same grounds. The field of LBD has not developed any such benchmark so far, even though this would bring the best compromise between the replication method (proper discoveries but very few instances) and time-sliced evaluation (large but very noisy sample).

This also raises the difficult question of the definition of a discovery. Clearly there is no simple and objective definition, but at least some general criteria can be established. In this work we define a discovery as a meaningful relation between two concepts, i.e. a relation which satisfies these two conditions:

- There is evidence in the literature that the relation is valid (this excludes cooccurrences which happen by chance);
- The relation must be insightful, i.e. relations which are trivially true are excluded.

We propose to interpret the concept of discovery as a spectrum which ranges from unambiguously positive cases to unambiguously negative cases. From this point of

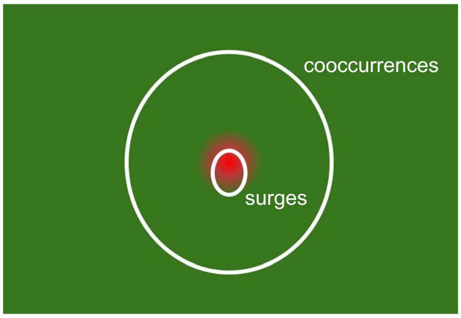

Fig. 1 Schematic diagram of the positive/negative discovery cases. The full area represents the set of all possible relations, i.e. the cartesian product of the indvidual concepts. The red area represents relations which clearly qualify as discoveries (positive cases) and the green area represents relatons which clearly do not (negative cases). The gradient represents uncertain cases which might be considered on either side. The small white elipsis represents the target set of relations and the large white circle represents all the cooccurrences

view, the few discoveries traditionally used by the *replication method* would belong to the small number of top positive cases, while the large set of cooccurrences used by the *time-sliced method* contains mostly negative cases. Our aim is to determine a measure of literature impact which reliably represents the "discovery level" of a relation across this spectrum. Such a measure would significantly improve the time-sliced evaluation method, since it would allow selecting a subset of the most positive relations as discoveries, instead of the noisy full set of cooccurrences. In other words, our method can be seen as a filtering step applied to the gold-standard set of discoveries considered by the time-sliced method: we propose substituting the large but low-quality gold-standard data with a smaller (but still large) number of high-quality relations in the time-sliced evaluation method, preserving every other part of the method.[2] Compared to the original time-sliced method, this would lead to a much better separation between the positive and negative discovery cases, as illustrated in Fig. 1. With this modification, the time-sliced evaluation would be more reliable thanks to the quality of the relations, and hopefully become more standard for LBD evaluation in the future.

Impactful discoveries

In this article we propose a method to build a benchmark dataset of discoveries based on the impact that a discovery has on the scientific literature.[3] The research ecosystem relies on community-based evaluation: through peer reviewing, journal reputation and citations, experts in a field evaluate each other's work and estimate the value of their contributions. In this perspective, it seems intuitively sound to rely on the prevalence of a relation in the literature as an indicator for the interest or importance of this relation. In particular, one expects a significant discovery to generate a large increase in the number of mentions of this relation in the following years, as the community studies this discovery and explores its implications. It is reasonably straightforward to measure the frequency of any relation from a corpus of the target literature, for example using

[2] In particular, one should still use the first year of cooccurrence as the cut-off year in order to avoid any data leakage.

[3] It is worth noting that alternative methods could be considered to mine discoveries, for example manual curation by experts, mining review articles, based on bibliometrics, etc.

the collection of biomedical abstracts provided by Medline.[4] Nonetheless the question of measuring *scientific impact* is not as simple as capturing a surge in frequency, because various factors other than a discovery event can affect frequency. Informally, the effect of a discovery on the literature can be described as a statistical outlier, i.e. an unusual event which stands out from the ordinary patterns.

It is worth noting that this approach assumes that the discovery is immediately recognized as such by the community and garners citations. Davies [15] explains that some discoveries can be premature or postmature, both cases resulting in a lower impact (or no impact at all) in the scientific community. Assuming that literature impact is used as a proxy for discoveries, these false negative cases are possible but arguably rare.

Another potential limitation of this approach is that it is unclear whether impactful discoveries are the ones that matter for LBD. Some important discoveries might have a very low number of cooccurrences, for example because they happen in a small research community. In fact, it can be argued that rare relations are more likely to be overlooked, and therefore more likely to lead to a LBD-based discovery. In this sense, a system which performs well on a dataset of impactful discoveries might not be good at finding other types of discoveries. Nevertheless, there is little doubt that impactful discoveries represent an important subset of discoveries, and could provide a decent proxy for the evaluation of LBD systems given the current limitations of the state-of-the-art.

Finally, it is useful to compare this approach to the one taken by Peng et al. [16]: in this paper the authors focused on the pairs of concepts which appear to be linked even though no article in the literature contains both of them. The authors identify such "gaps" by calculating the expected prevalence of every pair of concepts among those found individually in a set of related articles. The "gaps" are pairs which occur less often in the data than expected. Thus both this approach and ours try to detect unusual patterns in order to capture potential discoveries: Peng et al. [16] considers the absence of an expected cooccurrence as a potentially meaningful anomaly, whereas we are looking for anomalously high level of cooccurrences as a marker for impact. Many of the obstacles in both methods are common, in particular the need to filter out spurious relations (see in "Data and preprocessing" and "Post-processing" sections).

Method

Data and preprocessing

The experiments carried out in this paper use Medline as the source literature. Medline is a database containing more than 31 million references to articles published in life sciences journals.[5] Every reference in Medline is annotated with a set of MeSH descriptors[6] which represent the main biomedical concepts relevant to the article. We opt to use the Medline MeSH decriptors as concepts because these have been carefully annotated/reviewed by experts, thus the risk of error in the data is very low. However MeSH concepts are coarse and generic compared to other biomedical ontologies. The method

[4] https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

[5] The version of the data used in the experiments was downloaded in January 2021. The full code and instructions are provided at https://github.com/erwanm/medline-discoveries.

[6] https://meshb.nlm.nih.gov.

Table 1 Number of concepts and relations at different stages of filtering

Data	Unique concepts	Unique relations
Full data	29,406	47,291,526
Frequency ≥ 100	26,268	1,813,710
ND filtered	1803	290,797

described below can be applied to alternative representations of the literature, for example using PubTator Central[7] [17] or the UMLS Metathesaurus.[8] It could also be applied to richer semantic descriptions of the relations, e.g. DrugX-TREATS-DiseaseY, such as SemRep.[9,10] Also we consider only document-level cooccurrences since MeSH descriptors are provided by document, but the method could naturally be applied to sentence-level relations as well.

The year of publication and the MeSH descriptors are extracted for every entry using a modified version[11] of the "Knowledge Discovery" code by Lever et al. [14]. Every pair of MeSH descriptors in the same article defines a cooccurrence between two concepts. The raw data is processed using the "TDC Tools" repository[12] in order to obtain the frequency of (1) every concept (MeSH descriptor) by year and (2) every cooccurrence between two concepts by year. Additionally the pairs with less than 100 cooccurrences across all years are removed in order to prevent a large amount of noisy relations in the data.[13] The range of years is also filtered from 1950 onwards to avoid the low data volume of the early years.

Many of the examples and experiments presented below rely on a subset of the literature related to Neurodegenerative Diseases (NDs), the authors' primary domain of interest. The ND subset is obtained by first selecting all diseases which have the concept *Neurodegenerative Diseases* (D019636) as ancestor in the MeSH hierarchy, and then selecting all the concepts which cooccur at least once with any of these target concepts (this is equivalent to filtering all the articles which contain at least one of the target concepts and then selecting all the concepts in these documents). This filtering is purposefully loose in order for the dataset to include a broad range of concepts and relations with varying levels of specificity. The final dataset contains 291k distinct relations and 1.8k distinct concepts (see details in Table 1) with their frequency by year.

Measuring literature impact

In order to detect literature impact, a *trend* indicator is calculated for every year y and every relation (c_1, c_2). First, a moving average over a window of n years is applied in

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

[8] https://www.nlm.nih.gov/research/umls/index.html.

[9] https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html.

[10] Naturally such rich semantic relations are very useful for LBD [18]. But the extraction of the relations from the text is a potential source of error, this is why we adopted the most basic and reliable representation available.

[11] https://github.com/erwanm/knowledgediscovery.

[12] https://github.com/erwanm/tdc-tools.

[13] Rationale: relations with less than 100 occurrences are less likely to have an exceptionally high impact, i.e. to qualify as "impactful discoveries".

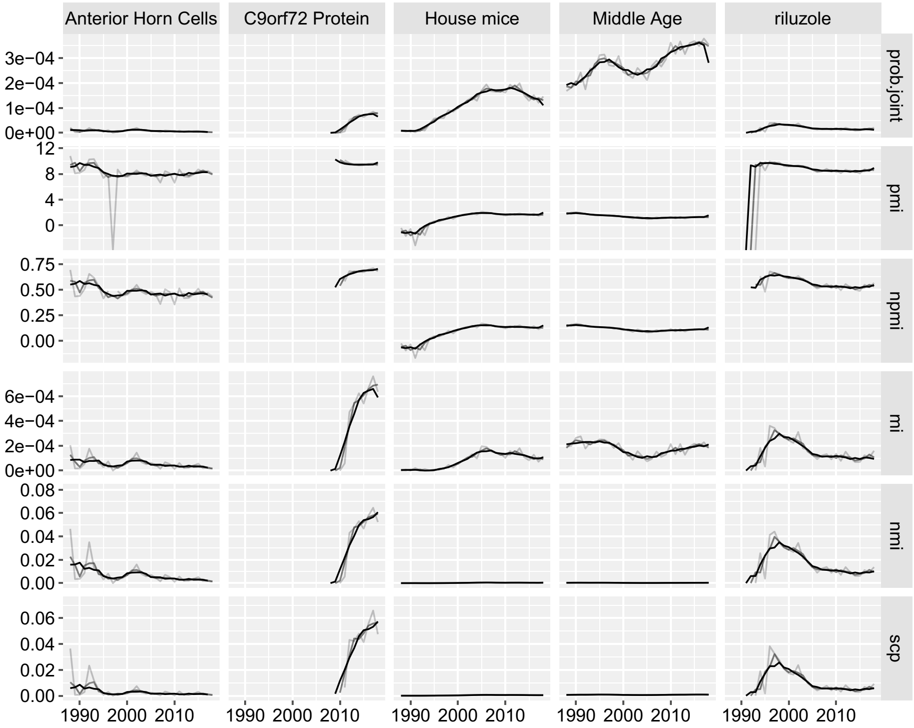

Fig. 2 Joint probability and other association measures in the period 1988–2018 for the concept *Amyotrophic Lateral Sclerosis* (D000690) paired with five concepts: *Anterior Horn Cells* (D000870), *C9orf72 Protein* (D000073885), *House mice* (D051379), *Middle Age* (D008875) and *Riluzole* (D019782). In one plot the different curves represent the values for different moving average window sizes, from the smallest size 1 (lightest) to the highest size 5 (darkest)

order to smoothen the frequency variations and capture the trend more accurately. Then a measure of statistical association is calculated for every year independently. Similarly to other LBD works [5, 6], we use a set of standard measures: Pointwise Mutual Information (PMI) as well as Normalized PMI (NPMI), the latter being less biased than PMI towards low frequency relations [19]; Mutual Information (MI) and Normalized MI, which take into account all the cases of concepts c_1 and c_2 appearing or not (we follow the definitions from [19]); and Symmetric conditional probability (SCP), the product of the two conditional probabilities. Although the joint probability is unlikely to be a good indicator, it is included as a baseline measure.[14]

Figure 2 shows the evolution between 1988 and 2018 of the joint probability and other measures between *Amyotrophic Lateral Sclerosis* (ALS) and five distinct concepts. These cases illustrate the diversity of the patterns captured by different measures. For example, the relations of ALS with concepts *house mice* or *middle age* have important changes across time in terms of probability (first row), but they do not, or less, in terms of statistical association (remaining rows). On the contrary, *C9orf72 Protein*[15]

[14] As opposed to using the raw frequency, the probability prevents the increase in number of publications (around 4% more every year) to artificially inflate the trend of a relation.

[15] The identification in 2009 of C9orf72 repeat expansions as the most common genetic variant associated with both ALS and frontotemporal dementia (FTD) confirmed the previously recognized pathobiologic association between the two conditions.

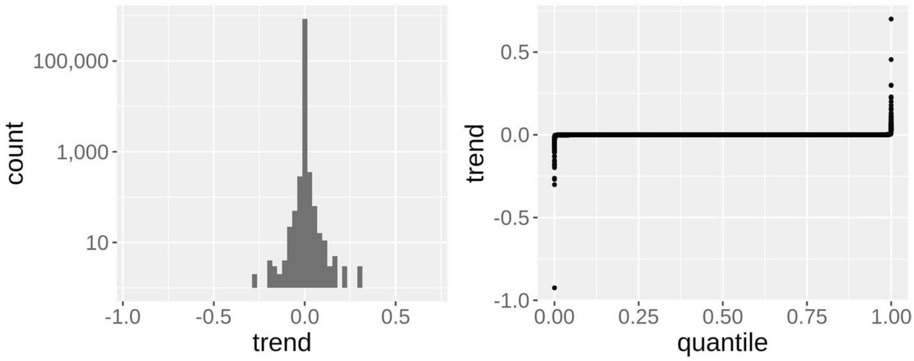

Fig. 3 Two views of the trend distribution for 20,000 random relations (NMI, window size 5, diff indicator). Left: histogram with logarithmic scale on the Y axis. Right: quantile plot (data points sorted by trend value on the X axis with their value shown on the Y axis)

and *riluzole*[16] have moderate frequency changes but show a strong surge according to some of the association measures. The relation of ALS with *anterior horn cells*[17] has a moderate evolution across time according to most measures.

The association measure is meant to represent the importance of a relation at any given time. The next step consists in measuring how this value evolves across time, i.e. its *trend* from 1 year to the next. Two simple indicators are considered:

- *diff* is the difference between the association values of the 2 years: $v_y - v_{y-1}$.
- *rate* is the relative rate of the association values of the 2 years: $\frac{v_y - v_{y-1}}{|v_{y-1}|}$.

Detecting surges

Since the trend indicator is a continuous numerical value, a threshold parameter is needed in order to separate regular years (negative, low or moderate trend value) from the years marked by a surge (high trend value). Naturally, the level at which a trend indicator becomes a surge is subjective and may depend on the target application.

In the case of building a test set of discoveries for LBD evaluation, the cases of interest are the most dramatic surges since one wants to maximize precision, i.e. to minimize false positive errors. Figure 3 shows two views of the trend distribution for the NMI measure with the *diff* indicator (other measures and indicators show the same general pattern). Even on a logarithmic scale, the regular histogram is not very informative due to the extremely high proportion of points close to zero. The quantile plot is more insightful, as it shows two clear inflection points at the extreme ends of the curve: while the vast majority of the points lie so close to zero that the curve looks flat, at both ends a few points get significantly farther from zero. The inflection point is determined by finding the point which maximizes the product of the normalized trend and the quantile (in other words, finding the largest rectangle starting from the top left corner of the quantile plot and having its bottom right corner on the curve).

[16] *Riluzole* is a drug used in the treatment of ALS; it was introduced in the US in 1995 and in the EU in 1996.

[17] A nerve cell in the spinal cord, rhombencephalon, or mesencephalon.

This method leads to selecting only the cases where the trend is exceptionally high, thus obtaining a very small proportion of pairs (relation, year) above the cut-off point: among the combinations of 6 measures, 2 trend indicators, and 3 sliding window sizes (1, 3 and 5), the median number of surges found represents only 0.2% of the pairs (relation, year) and 2.9% of the unique relations (cooccurrences). The majority of the combinations of parameters results in between 800 and 30,000 unique relations with surges, among a total of 290,797 possible relations (between 0.2 and 10%).

Post-processing

Several optional post-processing steps are also implemented, their use depends on the target application:

- The discovery year can be adjusted in cases where the surge is found in a year where the real frequency is zero. This may happen due to the moving average window, which occasionally creates high frequency values before the relation even exists. In such cases the surge year is shifted to the next year with non-zero frequency within the moving average window. Additionally, in the perspective of mining discoveries, the earliest surge is selected whenever a relation is found to have several years with a surge, i.e. any surge year later on is filtered out. However multiple surges in a relation can be meaningful in the context of some different application, for instance in the exploration of the relations which are "trending" at some particular point in time.
- Initial observations show that a large number of relations involve abstract and/or generic concepts, such as *Data Interpretation, Statistical* (D003627), *Cross-Over Studies* (D018592) or *Quality of life* (D011788). These relations may reflect real evolutions of the biomedical domain, especially methodological and technical innovation in biomedical research, but they are not primarily biomedical in nature and are often difficult to date precisely. This is why the UMLS Semantic Groups[18] can be used to filter the concepts types, by default selecting only the relations involving the four groups *'Anatomy'*, *'Chemicals and Drugs'*, *'Disorders'* and *'Genes & Molecular Sequences'*.
- A significant number of relations involve two concepts very closely related to each other, such as *Synucleins* (D051843) and *alpha-Synuclein* (D051844), *NF2 gene* (D016515) and *Neurofibromin 2* (D025581), *Presenilin-1* (D053764) and *Presenilin-2* (D053766). Naturally these relations tend to have simultaneous surges when their main concept has a surge within another relation since the two related concepts are frequently mentioned together. These spurious cases can also be filtered out by discarding relations in which at least one of the concepts has a high conditional probability with the other, indicating a trivial relation. The selection of the threshold is a trade-off between precision and recall: a high threshold means that the resulting list includes some trivial relations, whereas a low threshold eliminates some real discoveries from the list.

[18] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/documentation/SemanticTypesAndGroups.html.

Table 2 Gold standard dataset of 12 ND discoveries

MeSH 1	MeSH 2	Term 1	Term 2	Year
D002894	D006816	Chromosomes, human, pair 4	Huntington disease	1983
D000071617	D010300	Protein deglycase DJ-1	Parkinson disease	2003
D000690	D057180	ALS	Frontotemporal dementia	2009
D000073885	D057180	C9orf72 Protein	Frontotemporal dementia	2009
D000073885	D000690	C9orf72 protein	ALS	2009
D000072105	D000690	Superoxide dismutase 1	ALS	1993
D000068836	D000544	Rivastigmine	Alzheimer's disease	1997
D007980	D010300	Levodopa	Parkinson disease	1961
D007980	D020734	Levodopa	Parkinsonian disorders	1961
D000544	D001714	Alzheimer's disease	Bipolar disorder	2019
D004298	D010300	Dopamine	Parkinson disease	1971
D004298	D012559	Dopamine	Schizophrenia	1977

Results and analysis

Observations of the parameters

First, we analyze different aspects of the method using the ND subset (see in "Data and preprocessing" section). A small gold standard dataset of 12 ND discoveries is created in order to analyze the method (Table 2). This dataset is built using external sources to collect the year of discovery.[19] Reliably assessing the true year of discovery is a consistent obstacle in the evaluation of LBD. This issue sometimes causes errors, because if the discovery year is postdated then the performance of a LBD system may be overestimated [20]. The choice of the relations is arbitrary and is partly based on whether the relation has a clearly established year of discovery, therefore the dataset cannot be interpreted as a representative sample. Even in the case of discoveries which are clearly recognized in the ND field, there is often ambiguity about the exact date; for instance, the causal gene for Huntington's disease was approximately located in 1983 but precisely located on chromosome 4 in 1993.

This small-scale ND gold standard dataset is not meant as a reliable evaluation of the method, but as a way to compare the different parameters of the method (see in "Measuring literature impact" section). Additionally it can only be used to measure recall, i.e. the proportion of gold discoveries detected by the method. A discovery is considered as detected if the relation is retrieved and its surge year is within the window $y \pm n$, where y is the true discovery year and n is a fixed constant. Precision would require a full annotated subset and thus cannot be measured with this method.

The effect of the parameters of the method is evaluated by comparing their performance against the ND gold standard dataset. The surges are extracted for every configuration of parameters among: three window sizes for the moving average (1, 3 and 5); the six association measures (joint probability, PMI, NPMI, MI, NMI and SCP); and the two trend indicators (*diff* and *rate*).

[19] This criterion was established to ensure that the relations were uncontested discoveries. It turned out that this caused the rejection of many candidate relations, leaving only a few remaining discoveries. Nonetheless it was decided not to relax the conditions, i.e. to favour quality over quantity.

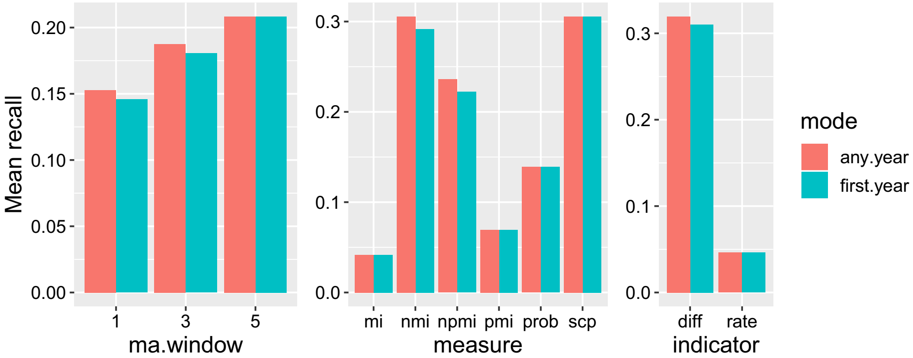

Fig. 4 Average recall within $y \pm 3$ years on the gold standard ND dataset (12 discoveries) by parameter. For each parameter value, the mean recall is calculated across the values of the other two parameters. The two colours represent whether all the surges are taken into account (*any year*) or the first year filter is applied (*first year*, see "Detecting surges" section)

For every parameter, Fig. 4 shows the average performance (recall) on the dataset, averaging across the values of the other two parameters. Performance increases with the window size of the moving average: 5 is better than 3, which is better than no moving average (1). The NMI and SCP measures perform best, followed by NPMI. PMI and MI perform poorly, more so than even joint probability. Finally the *diff* indicator performs drastically better than *rate*. Importantly, the *first year* filter (see in "Detecting surges" section) decreases performance only slightly compared to keeping all the surges; this is an indication that the method works as intended: for every relation which has several high surges, the first surge is very likely to be detected around the true time of discovery.

The best performing individual configurations are consistent with the results by parameters. SCP/diff/5 performs best, identifying the correct discovery year in 8 cases out of 12 (recall 0.67). It is followed by 6 configurations which perform equally well (7 cases out of 12; recall 0.58): NMI with window 1, 3 or 5, SCP with window 1 or 3 and NPMI with window 5 (all with the *diff* indicator). These results confirm the superiority of the NMI and SCP measures, but the small size of the dataset does not allow any fine-grained comparison between the configurations.

Discoveries across time

In this part we investigate the distribution of the surges predicted by the method across time. There are various potential biases related to the time dimension: for example, the volume of data is not uniform across time, since the number of entries in Medline increases by approximately 4% every year. There can also be artifacts due to the construction of Medline as a resource.[20] A distribution of the detected surges can be observed in Fig. 5 (middle); the distribution of the first cooccurrence year for all relations, i.e. the input data, is also shown for comparison (top). The surge patterns follow the first cooccurrence patterns quite closely, evidencing the absence of any visible bias due to the surge detection method. Most of the years have between 300 and 600 surges, which represents around 10% of the number of first cooccurrence relations. The peak in

[20] For example, the proportion of entries with an abstract in Medline jumps from 5% in 1974 to 40% in 1975.

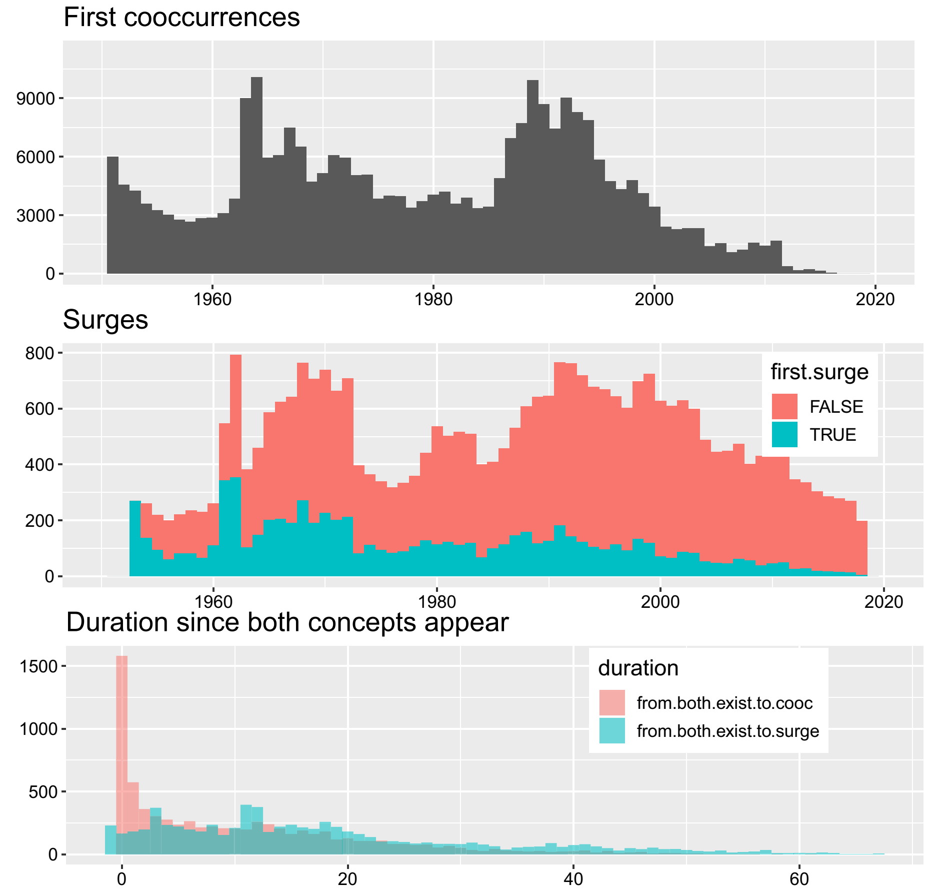

Fig. 5 Top: number of unique relations with frequency higher than 100 by year of first cooccurrence. Middle: number of surges by year, in blue for the earliest surge of a relation, red for subsequent surges. Bottom: difference between the earliest year where both concepts appear individually and (1) their first cooccurrence (red), and (2) their first detected surge (blue). Surges parameters: SCP measure, window 5, *diff* indicator

the 1960s is likely an artifact due to the construction of Medline. The decrease after the 2000s may be partly due to the filtering of the relations which have less than 100 cooccurrences (see in "Data and preprocessing" section), since relations which appeared in recent years had less time than older ones to accumulate 100 mentions. It is possible that the early years (50 s and 60 s) contain some spurious surges due to the low volume of data and the introduction in the data of concepts which might have existed before. For example, the relation *Adrenal Medulla* (D000313) and *Pheochromocytoma* (D010673) is detected as surging in 1952, although the discovery happened earlier. Nevertheless, other cases among the earliest surges detected appear to be valid, such as the relation between *adenosine triphosphate* (D000255) and *Stem cells* (D013234) which has a surge detected in 1952 [21, 22].

We also study how long after the introduction of the two concepts surges happen. Given a relation between two concepts c_1 and c_2 which appear for the first time at years y_1 and y_2 respectively, the earliest possible year for a cooccurrence (and consequently for a surge) is $max(y_1, y_2)$. The bottom plot of Fig. 5 shows the distribution of $y_c - max(y_1, y_2)$ and $y_s - max(y_1, y_2)$, with y_c the year of the first cooccurrence and y_s the year where the first surge is detected. With parameters *SCP/diff/5*, the first surge happens in average 7.6 years after the first cooccurrence. Among the relations which have a surge, 21% have their first cooccurrence the same year as the two concepts appear, whereas only 5.3% have their first surge this year. Similarly, 10% of the relations have a surge in the first 2 years while 30% cooccur in the first 2 years. Thus in a large number of cases, the first cooccurrence appears at the same time or soon after the first year where both concepts exist. However the first surge appears later most of the time, indirectly illustrating an important difference between the time-sliced method and our method: the former conflates cooccurrence and discovery, whereas the latter waits for evidence that the relation is an actual discovery.[21]

In most cases, the first surge occurs within 20 years of the two concepts appearing in the data. Nevertheless it is also possible for the surge to happen much later; in some cases, both concepts exist from the starting point of our dataset (1950) and have a first surge in the 2010s. For example, *Spinal Muscular Atrophy* (SMA; D009134) and *Oligonucleotides* (D009841) first appear in the data in 1951 and surge only in 2016, when a clinical trial for an antisense oligonucleotide drug for SMA proved successful [23]. However there are also questionable cases with some relations between two general concepts; for example, the relation between *Muscle Spasticity* (D009128) and *Motor Neuron Disease* (D016472) has its first detected surge in 2017, while both concepts exist in the data since 1950. This case likely corresponds to a "gap", as described by [16] (see "Impactful discoveries" section).[22]

Comparison against the time-sliced method

Finally, we compare our method against a baseline representing the time-sliced method (see "Motivations" section). In a real LBD evaluation setting, a cut-off year would be selected, the LBD system would be applied to the data before the cut-off year, and its predictions would be compared to the "true discoveries" happening after the cut-off year. The set of "true discoveries" is determined by the evaluation method: recall that the state-of-the-art time-sliced method consists in considering every cooccurrence of two terms as a "true discovery". Here the context is different, because we aim to compare the two evaluation methods themselves, not apply them in order to evaluate some LBD system. In this context, the two methods can be seen from the point of view of mining discoveries from the existing literature, where the time-sliced method acts as a simplistic baseline where every existing relation (two terms cooccurring) is automatically labelled as positive, as opposed to distinguishing relations which exhibit characteristics associated with a discovery (for example significant literature impact in our method). Thus the two evaluation methods can be compared simply by examining the set of relations they return, and determining which one is more likely to capture true discoveries.

[21] Similarly, our approach contrasts to [16]'s method in this perspective: [16]'s method searches for anomalously low cooccurrences in order to extract potential discoveries, therefore it virtually considers any new cooccurrence as a discovery and is sensitive to the introduction of new concepts in the literature. By contrast, our method seems able to handle this issue, thanks to fact that it doesn't rely on simple cooccurrences.

[22] The case of *Muscle Spasticity* (D009128) and *Motor Neuron Disease* (D016472) seems to belong to the "low hanging fruit" category in the authors' typology; this case is a surprisingly long gap for this category.

For our method we select specific values for the parameters, based on the previous analysis (*SCP/diff/5*). The baseline set of discoveries is obtained by extracting the N most frequent relations in the full dataset,[23] together with their first year of cooccurrence (see details in Table 1).[24] The original time-sliced method would normally include every cooccurrence which appears in the data: in our dataset (after relations with less than 100 cooccurrences were filtered out), this represents 108,794 unique relations after applying the post-processing steps. However N is chosen to be equal to the number of relations returned by our method ($N = 9092$), in order to make the two lists of relations comparable. The same post-processing steps are applied to both methods: filtering years 1990–2020, maximum conditional probability 0.6,[25] and filtering of the four groups *'Anatomy'*, *'Chemicals and Drugs'*, *'Disorders'* and *'Genes'*. The two resulting lists of discoveries are evaluated as follows: for each list, the top 100 relations are selected as well as a subset of 100 relations picked randomly in the list. Then the four subsets of 100 relations are randomly shuffled into a large dataset which is then annotated manually.[26] The final list contains no indication of which subset a relation comes from, so that the annotator cannot be influenced in any way. The annotation process is simplified in order to minimize the subjectivity involved in deciding whether a relation qualifies as a discovery or not. Every pair of concepts is labelled as one of three possibilities 'yes', 'no', 'maybe' regarding the discovery status. The annotator relies on Google Scholar queries with the two concept terms in order to determine their status:

- If the top results of the query show some evidence of a significant, non-trivial, new and impactful relation between the two concepts, then the relation is annotated as 'yes'. This requires at least one fairly clear title or abstract mentioning the relation as a discovery, with a healthy number of citations. The year of the main article is reported as gold-standard year (whether it is close to the predicted year or not).
- If there is evidence that the relation is either trivial, questionable or has very little impact (few papers or citations), then it is labelled as 'no'. This includes for example obvious relations, e.g. "Adrenergic Receptor—Adrenergic Antagonists", and relations involving trivial terms, e.g. "Traumatic Brain Injury—Neurons".
- In any other case, the status is considered ambiguous and the relation is labelled as 'maybe'. This includes cases where the annotator cannot understand the articles, has doubts about the originality, or the citation count is moderate. These cases are ignored in the evaluation results.

It is worth noting that this annotation policy is fairly strict regarding the discovery status of the relation: for example, in many cases a discovery exists with one of the

[23] i.e. not only the ND subset.

[24] We are grateful to an anonymous reviewer for suggesting this idea. The ranking by frequency is commonly used in LBD systems in order to show the most important relations first, therefore this is a reasonable baseline method.

[25] The conditional probability threshold is a very important parameter which controls the precision/recall tradeoff: here it is set to an arbitrary value of 0.6 (a balanced tradeoff in our tests), but a higher value would filter out more noise and result in a cleaner set of discoveries. Of course, the risk would be to remove some true discoveries then. This actually raises an important point: is it really meaningful to evaluate LBD using a "perfect" set of discoveries, by keeping only the most obvious ones? Not necessarily, because the "borderline" discovery cases could be even more relevant since they are potentially harder to find.

[26] A single annotator, one of the authors, carried out the full annotation process.

Table 3 Results of the SCP surges versus *time-sliced* baseline on 400 manually annotated relations

label	Subset: random 100		Subset: top 100	
	SCP/diff/5	**Baseline**	***SCP/diff/5***	**Baseline**
Yes	35	20	31	15
No	44	47	50	69
Maybe	21	33	19	16
Precision	44.3%	29.8%	38.3%	17.8%
χ^2 test yes/no (*p* value)	0.10426		0.00596	

The data was divided into 2 subsets for both methods: random selection of 100 relations (among the full 9092 relations), and top 100 relations (see details in "Comparison against the time-sliced method" section)

terms but the other term is only indirectly related; such cases would be labelled as negative. The non-trivial condition also discards many relations which could potentially qualify as discoveries. These strict criteria are intended to make the annotation process as deterministic as possible, even though real applications of LBD might consider a larger proportion of predicted discoveries as relevant. Thanks to this annotated dataset, it is possible to estimate the precision (proportion of true discoveries among the predicted ones) of our method and compare it against the baseline.

The results are presented in Table 3. It is striking that both methods have a large amount of non-discoveries (marked as "no") mixed with the actual discoveries (marked as "yes"). Surprisingly, both methods obtain a lower precision for the subset made of the top 100 relations (by frequency for the baseline, by trend for our method) than the subset made of 100 random relations, even though the top relations are expected to be more likely discoveries. This could be due in part to the high randomness and possible annotation errors, but in the case of the baseline many of the top relations are visible outliers: due to the particular severity of the pandemic, the most frequent relations involve *Covid19* with various other generic concepts, often not qualifying as a discovery (e.g. *Coronavirus* and *Emotional stress*).[27]

The precision obtained by our method is higher than the baseline by 14.5 points for the random 100 relations, and by 20.5 points for the top 100 relations. This confirms that our method based on measuring impact offers a better quality set of discoveries, even if it is still far from perfect. Nevertheless, the χ^2 test between the two methods, considering only the 'yes' and 'no' categories, is significant only for the top 100 subset; this is consistent with the fact that the difference between the precision values is lower for the random 100 subset. But even if our method obtains more than twice the precision of the baseline among the top 100 relations, it is disappointing that it does not reach more than 40%. This is a serious limitation that hopefully future improvements will alleviate.

[27] The fact that many spurious *Covid19* relations appear among the top for the baseline is caused by the frequency-based ranking. Nevertheless many of these spurious Covid19 relations also appear with our method, but these are mostly spread across the ranking due to the SCP association measure. In the case of our method, no obvious pattern explaining this surprising observation was identified; we suspect that this results from the combination of unusual frequency patterns in some relations together with the specifics of the association measure (here SCP).

Discussion

LBD evaluation

As explained in "Motivations" section, the time-sliced method proposed by [11, 13] is the only existing formal evaluation paradigm for LBD, and is often not even used. The proposed method improves the time-sliced method by filtering out a large proportion of mostly negative cases. As shown in "Comparison against the time-sliced method" section, our method obtains a higher precision than the baseline, knowing that this comparison does not take into account the 92% least frequent relations of the time-sliced method.

Our method uses literature impact as a marker for "true" discoveries, leading to considering only a very small subset of the cooccurrences as positive cases: in the experiment presented above, the median number of relations above the surge threshold represents 3% of all the input relations (see "Detecting surges" section). Thus our method allows a more accurate balance between the positive and negative cases:

- In the time-sliced method, the cases considered as negative are perfect (since there cannot be a discovery without cooccurrence), but the vast majority of the cases considered as positive are actually negative.
- In our method, both the cases considered as negative and positive are slightly imperfect, but the latter are likely to represent true discoveries.

However, our method is only as good as its underlying assumption that literature impact is a reliable marker for discoveries. In this respect, experimental results in "Comparison against the time-sliced method" section are disappointing: the relations retrieved by our method are guaranteed by construction to have had an exceptional impact in the literature, in the form of a very strong surge as represented by the selected association measure. Nevertheless, the resulting relations still contain a large amount of noise. This might be related to the use of MeSH descriptors to represent the literature; future work using more advanced representations might obtain better results.

LBD and the time dimension

Traditionally, LBD models consider a static representation of the state of knowledge at a particular point in time. The time dimension is only taken into account for the purpose of evaluation. This implies that the existing relations in the dataset are considered equivalently likely to lead to a future discovery. Clearly, this is also a simplification: the scientific knowledge ecosystem follows a dynamic evolution, with some topics trending and others being abandoned (see [24]). The identification of a significant new relation often causes a thrust of research, simply because the new relation is more likely to help uncover more new knowledge than any older relation. It seems intuitively relevant for LBD methods to integrate the time dimension as well. For example, the ABC model relies on linking concepts in common between two heterogeneous subsets of the literature and drawing new relations by transitivity; it might be relevant as well to prioritize relations which were discovered at different times. More advanced LBD models might even be able to model the dynamic evolution of the topics and concepts across time in order to actually predict future trends, instead of only exploring the search space of all possible relations.

Exploration tool and other applications

As a result of this work, two datasets of surges detected from Medline are made available: the ND subset (see in "Data and preprocessing" section) and the surges obtained on the full Medline data.[28] A simple exploration tool is also made available.[29] This tool lets a user observe the relations retrieved by our method with various filtering options. For example, a user may search for recent discoveries related to treating a particular disorder by selecting the appropriate target concept, range of years and semantic group. This kind of tool offers a very synthetic view on the body of knowledge represented in the biomedical literature. More generally, the availability of a dataset of time-stamped discoveries may lead to novel applications and usages, for example as a methodological resource for systematic reviews.

Conclusion and future work

This work proposes to exploit the evolution of the literature across time in order to retrieve the relations which had the most significant impact in the past. We intentionally adopt a simple heuristic method based on descriptive statistics because this way the results are straightforwardly interpretable: it is guaranteed by construction that the extracted relations had an exceptional impact in the literature, in the form of a very strong surge as represented by the selected association measure. While experimental results show that there are still various issues even in this case, we think that this kind of interpretability is crucial in order to use the dataset as a benchmark: this way any potential user knows what kind of discoveries are included and the limitations of the dataset.

Compared to previous work in the evaluation of LBD, this approach can be used to replace the set of cooccurrences used as gold standard by the time-sliced method, hence offering a more meaningful way to evaluate LBD systems on a large set of discoveries. It might also pave the way for more fine-grained LBD methods, which could exploit these past discoveries to train supervised models.

Although simpler than LBD, the task of extracting impactful discoveries from the existing literature is not trivial. To some extent, the ability to reliably identify past discoveries can be seen as a prerequisite for the harder case of LBD, since the latter also needs to identify discoveries but with less information available as input. In particular, the distinction between meaningful and non-meaningful relations is complex. Existing LBD methods rely on simplifications which tend to mask this difficulty, but this question must be addressed if LBD is to become a well grounded methodological tool for the biomedical domain.

Finally, using automatically generated data for evaluating LBD methods is far from ideal. But given the current lack of good options for LBD evaluation, we believe that such an automatic method, even with important limitations, is a reasonable approach to

[28] Available at https://zenodo.org/record/5888572.

[29] https://brainmend.adaptcentre.ie/.

evaluate some LBD applications. Hopefully this can contribute to inspire new and better evaluation methods in the future.

Acknowledgements
The authors are grateful to the anonymous reviewers for their valuable comments.

Author contributions
EM conducted the analysis, designed the experiments and wrote most of the article. OH and MH contributed the initial ND dataset and interpreted the biomedical concepts and relations. DO'S contributed to writing the article.

Funding
This work was conducted with the financial support of Irish Health Research Board (HRB) BRAINMend project (Grant Number HRB-JPI-JPND-2017-1) and with the support of the Science Foundation Ireland Research Centre ADAPT (Grant Number 13/RC/2106_P2).

Availability of data and materials
The source data used in this article have been extracted from PubMed/Medline in 2021: https://pubmed.ncbi.nlm.nih.gov/download/. The implementation is published under open-source license and available at https://github.com/erwanm/medline-discoveries. A detailed documentation is also provided (https://erwanm.github.io/medline-discoveries/). The dataset resulting from the application of the system to the 2021 Medline data is published at https://zenodo.org/record/5888572. An online exploration tool is also provided at https://brainmend.adaptcentre.ie/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Cmpeting interests
The authors declare no competing interests.

Received: 20 February 2023 Accepted: 23 July 2024
Published online: 16 September 2024

References

1. Henry S, McInnes BT. Literature based discovery: models, methods, and trends. J Biomed Inform. 2017;74:20–32.
2. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7–18.
3. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med. 1988;31(4):526–57.
4. Thilakaratne M, Falkner K, Atapattu T. A systematic review on literature-based discovery workflow. PeerJ Comput Sci. 2019;5:235.
5. Pyysalo S, Baker S, Ali I, Haselwimmer S, Shah T, Young A, Guo Y, Högberg J, Stenius U, Narita M, et al. LION LBD: a literature-based discovery system for cancer biology. Bioinformatics. 2019;35(9):1553–61.
6. Crichton G, Baker S, Guo Y, Korhonen A. Neural networks for open and closed literature-based discovery. PLoS ONE. 2020;15(5):0232891.
7. Swanson DR. Undiscovered public knowledge. Libr Q. 1986;56(2):103–18.
8. Smalheiser NR. Literature-based discovery: beyond the ABCs. J Am Soc Inf Sci Technol. 2012;63(2):218–24.
9. Kastrin A, Hristovski D. Scientometric analysis and knowledge mapping of literature-based discovery (1986–2020). Scientometrics. 2021;126(2):1415–51.
10. Ganiz MC, Pottenger WM, Janneck CD. Recent advances in literature based discovery. Technical report, Lehih University; 2005.
11. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. J Biomed Inform. 2009;42(4):633–43.
12. Yetisgen-Yildiz M, Pratt W. Evaluation of literature-based discovery systems. In: Bruza P, Weeber M, editors. Literature-based discovery. Berlin, Heidelberg: Springer; 2008. p. 101–13.
13. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Stud Health Technol Inform. 2001;84:1344–8.
14. Lever J, Gakkhar S, Gottlieb M, Rashnavadi T, Lin S, Siu C, Smith M, Jones M, Krzywinski M, Jones SJ. A collaborative filtering-based approach to biomedical knowledge discovery. Bioinformatics. 2017;34(4):652–9.
15. Davies R. The creation of new knowledge by information retrieval and classification. J Doc. 1989;45(4):273–301.
16. Peng Y, Bonifield G, Smalheiser NR. Gaps within the biomedical literature: Initial characterization and assessment of strategies for discovery. Front Res Metr Anal. 2017;2:3.
17. Wei C-H, Allot A, Leaman R, Lu Z. Pubtator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1):587–93.
18. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. In: AMIA annual symposium proceedings, vol 2006. American Medical Informatics Association; 2006, p. 349.
19. Bouma G. Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of GSCL. 2009, p. 31–40.

20. Kostoff RN. Validating discovery in literature-based discovery. J Biomed Inform. 2007;40(4):448–50. https://doi.org/10.1016/j.jbi.2007.05.001.
21. Bendall J. Effect of the 'marsh factor' on the shortening of muscle fibre models in the presence of adenosine triphosphate. Nature. 1952;170(4338):1058–60.
22. Bendall J. A factor modifying the shortening response of muscle fibre bundles to ATP. Proc R Soc Lond Ser B Biol Sci. 1952;139(897):523–5.
23. Finkel RS, Chiriboga CA, Vajsar J, Day JW, Montes J, De Vivo DC, Yamashita M, Rigo F, Hung G, Schneider E, et al. Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study. Lancet. 2016;388(10063):3017–26.
24. Kastrin A, Hristovski D. Disentangling the evolution of medline bibliographic database: a complex network perspective. J Biomed Inform. 2019;89:101–13.

Publisher's Note