Health and Technology (2021) 11:1205–1217
https://doi.org/10.1007/s12553-021-00605-y

REVIEW PAPER

Literature-based discovery approaches for evidence-based healthcare: a systematic review

Sudha Cheerkoot-Jalim[1] · **Kavi Kumar Khedo**[2]

Received: 3 May 2021 / Accepted: 28 September 2021 / Published online: 25 October 2021
© IUPESM and Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Purpose Literature-Based Discovery (LBD) is a text mining technique used to generate novel hypotheses from vast amounts of literature sources, by identifying links between concepts from disparate sources. One of the main areas where it has been predominantly applied is the healthcare domain, whereby promising results, in the form of novel hypotheses, have been reported. The purpose of this work was to conduct a systematic literature review of recent publications on LBD in the healthcare domain in order to assess the trends in the approaches used and to identify issues and challenges for such systems.
Methods The review was conducted following the principles of the Kitchenham method. The selected studies have been scrutinized and the derived findings have been reported following the PRISMA guidelines.
Results The review results reveal useful information regarding the application areas, the data sources considered, the approaches used, the performance in terms of accuracy and reliability and future research challenges. The results of this review will be beneficial to LBD researchers and other stakeholders in the healthcare domain, by providing them with useful insights on the approaches to adopt, data sources to consider, evaluation model to use and challenges to reflect on.
Conclusion The synthesis of the results of this work has shed light on recent issues and challenges that drive new LBD models and provides avenues for their application in other diverse areas in the healthcare domain. To the best of our knowledge, no such recent review has been conducted.

Keywords Literature-based discovery · Evidence-based healthcare · Knowledge translation · Systematic review

1 Introduction

Healthcare management, being one of the highest priorities of most governments, attracts huge investments in terms of health and medical research worldwide. Medical research was found to be the main contributing factor in the improvement of health and longevity of individuals and populations in developed countries [1]. Researchers in the field are making new discoveries and generating knowledge, which has the potential to enhance healthcare delivery, improve patient health outcomes and reduce healthcare costs, thus strengthening the overall healthcare system and economy. This is only achievable if the knowledge is actually put into action [2]. However, the transfer of research findings into healthcare practice in the clinical setting, known as knowledge translation [3], is a very complex and slow process, often resulting in patients not being provided with the most appropriate care, although better treatment recommendations have been proposed and demonstrated. A frequently stated average time lag for knowledge translation is 17 years [4]. Understanding the various stages of knowledge translation and speeding up the process is a policy priority for many health research systems [4].

In order to leverage new medical research findings more quickly for the benefit of patients, medical practitioners are encouraged to adopt the practice of evidence-based medicine, whereby medical practitioners are expected to scrutinize the scientific and clinical research literature in their respective areas in an attempt to translate health research knowledge into effective healthcare action more quickly. However, due to the large volumes of biomedical literature available and the time constraints of medical practitioners, the practice of evidence-based medicine has become a major challenge [5].

✉ Sudha Cheerkoot-Jalim
s.cheerkoot@uom.ac.mu

[1] Department of Information and Communication Technologies, University of Mauritius, Reduit, Mauritius

[2] Department of Digital Technologies, University of Mauritius, Reduit, Mauritius

This limitation can be considerably overcome by the use of appropriate computation techniques for the automated or semi-automated knowledge extraction from relevant research literature. A broad term commonly used for such techniques is literature based discovery (LBD), whose main goal is to generate novel hypotheses from the vast available biomedical literature by discovering unknown associations in existing knowledge [6]. Recent advances in machine learning, text mining and statistical analysis techniques have spurred research in this field and have resulted in many publications on the design and application of LBD systems for various use cases in the biomedical and healthcare domains.

The purpose of this work is to perform a systematic literature review of recently published research papers on the application of LBD for evidence-based healthcare, with the objective of identifying and integrating the findings of the most relevant individual studies. It is expected that the results of this review will give insights on the different LBD approaches and tools used in various application areas in the healthcare domain. It will help establish to what extent research has progressed in the field, with a focus on performance criteria like effectiveness, accuracy and reliability. A main outcome would be to identify research challenges, which will invoke further studies and thus, provide avenues for future research in other areas in the healthcare domain. The Kitchenham guidelines for performing systematic literature reviews [7] was adopted and the reporting of this paper follows PRISMA (preferred reporting items for systematic reviews and meta-analysis) guidelines [8]. To the best of our knowledge, no such recent review has been performed for evidence-based healthcare.

1.1 Evidence-based healthcare

The challenges of knowledge translation have become a major concern to individuals who seek and need healthcare, healthcare providers, policy makers and funders of health services. The incorporation of scientific medical discoveries into practice guidelines and policies in the clinical setting can greatly improve healthcare delivery and patient health outcomes, and is the basis of evidence-based healthcare [9]. Evidence-based practice involves clinical decision making which considers the best and most up-to-date available scientific evidence, together with patient values and preferences, the clinical judgment of the medical practitioner and the context in which the care is provided [10]. Healthcare professionals seek evidence to support and justify any activity or intervention for patient care.

1.2 Literature based discovery in healthcare

In their practice of evidence-based medicine, medical practitioners are expected to scrutinize the best available evidence for making decisions about the care of individual patients. However, with the increasing volume of academic research papers and related structured knowledge resulting from medical research worldwide, they only focus on publications that are directly relevant to their respective area of specialization and often skip other potentially relevant research. Thus, discoveries in one field remain unknown to others and potential connections between sub-fields are often missed out [11]. This limitation can be greatly curbed by LBD, which can automate or semi-automate the analysis of online resources from disparate sources to find new discoveries. With the exponential growth of scientific literature, LBD is becoming an increasingly important tool for facilitating research [12].

LBD generates discoveries not yet published anywhere, by combining knowledge extracted from varied literature sources and therefore, supports hypothesis generation [13]. There are two modes of discovery in LBD, namely open discovery and closed discovery. Open discovery starts with a concept X and tries to generate a potential association between X and another concept Z, based on an intermediate concept Y. This follows from the ABC co-occurrence model, which states that if A and B are often associated to each other, and B and C are also often associated to each other, there may potentially be an association between A and C, even if this association is not mentioned in any research paper [14]. In contrast, in closed discovery, both the start concept X and end concept Z are known, and an association between X and Z is predicted, based on a hypothesis about the relationship between X and Z. This technique then attempts to demonstrate the hypothesis through an intermediate concept Y.

LBD approaches in healthcare are becoming essential, since biomedical knowledge is spread out across a larger number of publications [15]. Potential discoveries in healthcare can be associations that exist between biomedical concepts, which are not usually discussed together in the literature. Appropriate implementation of LBD techniques have the potential to predict future strong associations between these concepts [15] and therefore entails further research. In the LBD approach the starting concept X may be a disease and the end concept Z may be a treatment or cause for the disease. The results of such discoveries need to be further investigated through experimental methods or clinical studies.

2 Materials and methods

This review has been performed following the guidelines on undertaking systematic literature reviews by Kitchenham and Charters [7] and the reporting follows the PRISMA guidelines [8]. The methodology consisted of first setting out the research questions to give a focus for this review, followed by the specification of the search strategy, the application of assessment criteria for the selection of papers and finally the data analysis and extraction.

2.1 Research questions

Based on the objectives of this review, the research questions have been set out and elaborated as follows:

RQ1: What are the main application areas of literature based discovery in evidence-based healthcare? We seek to find out the different application areas in which the application of LBD techniques has proved to be successful in the healthcare domain.

RQ2: Which important/impactful literature sources are considered by researchers/practitioners for literature based discovery? The foundation of LBD is the large amount of scientific literature available for a specific field of study. It is therefore important to identify the different literature sources which have been harnessed for LBD in the different studies.

RQ3: Which specific literature based discovery approaches and tools have proven to be effective in the healthcare domain? Due to the peculiarity of the healthcare domain, LBD techniques have to be adapted to specific application areas. There is therefore the need to investigate the specific LBD techniques/approaches which are more relevant and effective for the healthcare domain.

RQ4: How do literature based discovery systems in the healthcare domain perform in terms of accuracy and reliability? Accuracy and reliability are imperative evaluation criteria for any computational technique in the healthcare domain, since a wrong intervention can lead to harmful consequences for the patient. We therefore study the different evaluation strategies used for LBD systems and find out their performance in terms of accuracy and reliability.

2.2 Search strategy and study selection

The search strategy involved the identification of potential research papers to be included in the review by performing a search on Google Scholar, with keywords '"Literature-based discovery" in health'. Google Scholar was chosen since it indexes scientific articles from various scholarly publishers and professional societies like Springer, ScienceDirect, ACM, IEEE Xplore, ResearchGate amongst others [16]. It also indexes biomedical-specific journals like the Journal of Biomedical Informatics, PLOS ONE and BioMed Central (BMC). Gusenbauer [17] performed a comparative study of academic search engines in 2019 and concluded that "Google Scholar is currently the most comprehensive academic search engine". Keyword search was then followed by a manual screening of reference lists of relevant primary studies to extend the search space.

2.3 Eligibility criteria

Based on the objectives of this systematic review, we have set some inclusion and exclusion criteria to guide the study selection process, as follows. The focus of this review being on recent advances in LBD techniques and approaches, we considered studies carried out during the last five years, that is, since 2015. We only considered peer-reviewed papers published in the English language. Primary studies were included while secondary and tertiary studies, like surveys, systematic reviews and meta analyses were excluded. During an initial screening of studies, we came across papers which describe general LBD techniques without showing their application in the healthcare domain. Such studies were not included, since the objective of this review was to get insights on the different approaches which are more appropriate for specific application areas of LBD. We thus considered papers which describe the use of LBD approaches in a specific application area in the healthcare domain.

The database search was performed on 2nd February 2021. The keyword search returned 650 results, after applying the filter on year of publication. The manual screening of reference lists of relevant studies returned 12 eligible studies. 8 duplicate studies were identified from the two sources, resulting in 654 studies to screen. After a rigorous screening of the titles and abstracts based on the inclusion and exclusion criteria, 29 studies were pre-selected for the review.

2.4 Quality assessment

After initial screening based on the inclusion and exclusion criteria, the pre-selected studies were assessed for "quality" in order to integrate more detailed inclusion and exclusion criteria. Based on the research questions, four quality assessment criteria were set as shown in Table 1. The possible outcomes for each criteria were "Yes" if the paper met the criteria and "No" if it did not meet the criteria. Two of the quality assessment criteria also had a "Partially" outcome.

During the quality assessment phase, appropriate scores were given to each pre-selected study. A score of 1 was given for a "Yes" outcome, 0 for a "No" outcome and 0.5 for a "Partially" outcome. Studies which obtained a score of at least 2.5 were included in the final review. This would allow for one "No" and one "Partially" outcome in the outmost scenario. After the quality assessment phase, 23 studies have been selected for the final review, based on the scores obtained. Figure 1 shows the PRISMA flow diagram for the study selection process.

Table 1 Quality Assessment Criteria

No	Quality Criteria	Outcome
QC1	Has the LBD approach used been described in detail?	Yes: The LBD approach used has been described in detail Partially: The LBD approach used has been briefly described No: The LBD approach used has not been described
QC2	Was there a discovery following the research work?	Yes: There was a discovery No: No discovery was made
QC3	Did the study include a concise evaluation strategy?	Yes: A concise evaluation was done Partially: The evaluation was not intensive No: No evaluation was done
QC4	Does the study give insights on research challenges and future directions?	Yes: The study gives insights on research challenges and future directions No: The study does not give insights on research challenges and future directions

3 Results

The selected studies were thoroughly analyzed with an objective to extract information which would give insights to the research questions. More particularly, the information extracted were: the medical application area in which LBD was utilized and the discovery made as a result of LBD, the literature source/s considered, the type of discovery (open or closed), the techniques and tools used in the LBD approach, the performance of the system and the challenges identified by the authors. The data synthesis is shown in Table 2.

4 Discussion

The selected studies were scrutinized with a major focus on the objectives of this review. The work of the various authors and their findings were mapped to the research questions and are discussed in the following sub-sections.

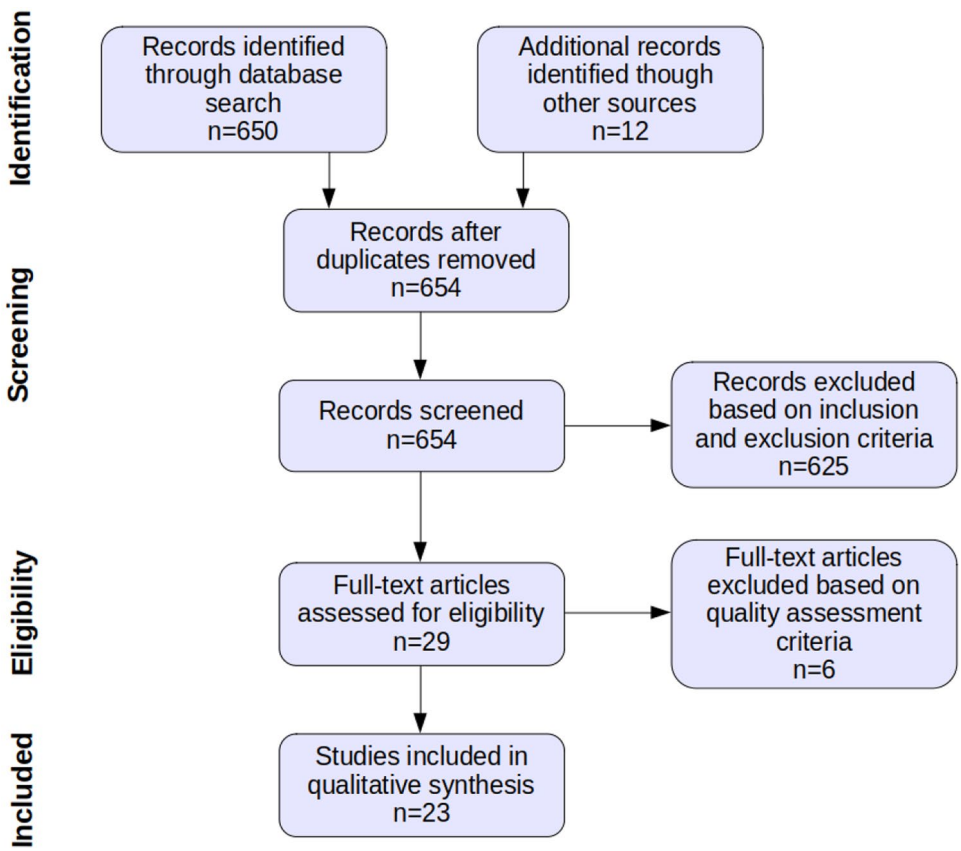

Fig. 1 PRISMA Flow Diagram for the Study Selection Process

4.1 RQ1: What are the main application areas of literature based discovery in evidence-based healthcare?

From the studies analyzed, it was found that LBD techniques have been implemented in a myriad of application areas in the healthcare domain, as described below.

4.1.1 Drug repurposing

Drug repurposing is one main application area in which researchers have put efforts, mostly because of the promising results achieved by the different LBD approaches proposed. Due to the huge costs and excessive amount of time involved in developing new drugs, it is regarded as a better alternative. Several studies [18, 19, 21, 23, 25] generated a list of potential drug-disease pairs by using drug-gene and gene-disease semantic predications. Phenotypes and symptoms have also been used as the linking concept between drug and disease [16]. Some studies have used knowledge-graph based drug discovery methods [18–20].

4.1.2 Pharmacovigilance and drug interactions

Pharmacovigilance involves the continuous monitoring of drug safety after drugs are put on the market, which is necessary since some adverse drug events (ADEs) remain undetected during clinical trials and unreported in adverse event reporting systems such as FAERS (FDA Adverse Event Reporting System). The health hazards that ADEs may pose to individuals motivate the extensive work on the application of various computational methods for pharmacovigilance. Authors of this study have either used an open LBD [15, 23, 24] or a closed LBD [22, 25] approach for the detection of drug/ADE pairs.

4.1.3 Identification of potential causes, therapies or treatments for specific diseases

LBD's potential to contribute to the advancement of the medical field has been demonstrated by the development of text mining systems which have been able to identify possible causes, therapies or treatments for specific diseases. Discoveries about connections between diet and degenerative diseases [34, 35] were made from scientific literature to support better understanding and treatment of such diseases. LBD techniques have been used for rehabilitation therapy repositioning for stroke [31] and treatment repurposing for inflammatory bowel disease [36]. Other discoveries were made in the area of cancer [32] and chronic kidney disease [33].

4.1.4 Explanation for the correlation between diseases

Disease comorbidity is very common and is a popular area of research in the medical community, because of its impact on the treatment of diseases. Knowledge of the association between diseases can significantly improve the understanding of the mechanisms of diseases, thus aiding in better prevention and treatment [37]. Thus, Chen et al. [37] have used an open LBD approach for the detection of associations among complex diseases. Closed LBD approach was also used for the explanation of the correlation between epilepsy and inflammatory bowel disease [38], and between myocardial infarction and depression [39]. Rather et al. [40] proposed the use of deep learning for the discovery of potential new biomedical knowledge.

Table 3 summarizes the main application areas and the number of studies for each.

4.2 RQ2: Which important/impactful literature sources are considered by researchers / practitioners for literature based discovery?

The main literature sources for LBD leveraged by authors of studies in this review are Medline (10 studies) and PubMed (13 studies). Medline, the bibliographic database of the US National Library of Medicine, is indexed with Medical Subject Headings (MeSH) terms, making search in the biomedical domain more effective. This explains its popularity among LBD researchers. PubMed on the other hand, is an interface to search Medline together with other additional biomedical content. Tools which extract data from PubMed and Medline, like Global Network of Biomedical relationships (GNBR) [20] and Semantic Medline Database (SemMedDB) [21, 26] have also been proposed. Apart from PubMed, Zhang et al. [25] also extracted data from CORD-19 (Covid-19 Open Research Dataset), which contains Covid-19-related literature, which may not yet be available on PubMed. An additional literature source, Chinese Science Database (CNKI), was used to extract herb-disease pairs in Traditional Chinese Medicine [26].

4.3 RQ3: Which specific literature based discovery approaches and tools have proven to be effective in the healthcare domain?

Since the data sources mainly consist of free-text, the main techniques behind LBD are text mining and natural language processing (NLP). Most LBD approaches proposed have extracted meanings from biomedical text by using Unified Medical Language System (UMLS) concepts and MeSH terms. The approaches used by authors of studies in this review are broadly categorized and described below.

Table 2 Data Synthesis

Study	Area of application / Discovery	Literature Source/s	Type of Discovery	Techniques Used	Tools Used	Performance Issues	Challenges
Rastegar-Mojarad et al. [18]	Drug repositioning – generation of potential drug-disease pairs	Medline Abstracts	Open	ABC Model of LBD, Semantic Predications, NLP	SemRep	Actual performance of the system could not be accurately benchmarked, Inherent limitation to evaluate confidence levels of generated pairs	Ranking of the generated candidates, Rigorous validation is required before proceeding to laboratory or clinical investigations
Yang et al. [19]	Identification of new candidates for repurposing as anticancer drugs	Medline Database	Open	ABC model of LBD, relationship extraction, text mining-based ranking method	Apache Lucene search engine	Higher precision can be achieved by the use of a comprehensive lexicon, Negative relationships not considered	Consideration of aliases, Normalization of gene and disease targets
Raja et al. [20]	Repurposing drugs for four diseases	PubMed Abstracts	Open	ABC model of LBD, Co-occurrence	KinderMiner	Evaluation was done by comparing the prediction score of annotated drugs and new drugs	Differentiation between positive and negative associations
Rastegar-Mojarad et al. [21]	Discovering drug-disease relations (drug repositioning and adverse drug reactions)	PubMed	Open	Classification of drug-disease relations into desired classes for ranking hypotheses	SemRep	Evaluation was done for balanced data sets only	Train and evaluate classifier using unbalanced data sets, Identification and removal of false positive candidates
Zhao et al. [22]	Discovery of potential drugs for diseases	PubMed	Open	Convolutional Neural Network, Logistic Regression, Attention mechanism, Path ranking algorithm	SemRep	Weak generalization, due to small data set used to train the model, A larger knowledge base may affect efficiency	Use larger data set by combining other drug-disease databases, Improve the NLP technology to be able to cope with a larger knowledge base
Sang et al. [23]	Discovery of candidate drugs for diseases	PubMed abstracts	Open	Logistic regression model	MetaMap, SemRep	Accuracy of MetaMap reduces because of inability to resolve word sense disambiguation, Considerable number of false predications	Development of high-quality NLP tools, Graph embedding to obtain long paths
Sosa et al. [24]	Drug repurposing for rare diseases	GNBR (Global Network of Biological Relationships) (PubMed abstracts)	Open	Knowledge graph embedding	N/A	Performance decreases, since only 'treatment' relationships are chosen, while potential other relationships are discarded	Failure to capture complex and indirect relationships.

Table 2 (continued)

Study	Area of application / Discovery	Literature Source/s	Type of Discovery	Techniques Used	Tools Used	Performance Issues	Challenges
Zhang et al. [25]	Drug repurposing for Covid-19	PubMed, CORD-19	Open	Knowledge Graph Completion, Translational and semantic matching models	SemRep, MetaMap	Accuracy was affected due to loss of information by the use of sub-graphs and the precision and recall of SemRep	Inclusion of other types of biological data
Xie et al. [26]	Identification of alternative herbs for drugs that cause side effects	PubMed, Chinese Science Database (CNKI)	Open	ABC model of LBD, Co-occurrence, Gene enrichment analysis	N/A	N/A	Inclusion of similarity of chemical compounds
Zhang et al. [27]	Exploration of interactions between cancer drugs and dietary supplements	Medline	Closed	Concept mapping, Machine learning-based filtering	SemRep, MetaMap	Limited precision and recall, Knowledge source shortcomings and linguistic issues	Use of machine learning for the automatic filtering of semantic predications
Malec et al. [28]	Pharmacovigilance – Detection of drug-ADE associations	Medline	Open	Feature selection, Co-occurrence based analyses	SemRep	Search for cofounders was relatively shallow, Reference data sets not perfectly accurate	Consideration of comorbidities and co-medications, Analysis of FAERS data instead of only EHR data
Mower et al. [29]	Prediction for unseen drug-side effect pairs	Medline citations	Open	Machine Learning, Composite feature vectors	SemRep, SIDER	Robust performance for the ESP-based model	Terminological mapping, Abstraction methodologies, Integration of observational data sources
Hristovski et al. [30]	Provide pharmacological and pharmacogenomic explanations for reported ADEs	Medline	Closed	Semantic relation extraction	SemBT, SemRep	N/A	N/A
Meng et al. [31]	Identification of possible rehabilitation therapies for stroke	PubMed	Open	ABC Model of LBD, Co-occurrence	E-Utilities	N/A	N/A
Pyysalo et al. [32]	Discoveries on the molecular biology of cancer	PubMed	Open and Closed	Machine Learning, Natural Language Processing, Co-occurrence based metrics	PubTator, Hallmarks of Cancer Classifier	System recognizes a single target response for each case and manual analysis is required	Inclusion of full texts
Kostoff and Patel. [33]	Identification of foundational causes of chronic kidney disease, Identification of treatments	Medline	Open	Co-occurrence	Vantage Point software	Uncertainty in the 'degree' of cause removal	The magnitude of the associations could not be determined, Identification of 'mix' of causes for an individual patient

Table 2 (continued)

Study	Area of application / Discovery	Literature Source/s	Type of Discovery	Techniques Used	Tools Used	Performance Issues	Challenges
Gubiani et al. [34]	Discoveries about connections between diet and degenerative diseases	PubMed	Open	Ontologies, RaJoLink	OntoGen	N/A	Validation of robust in-silico tools
Gubiani et al. [35]	Identify molecular links between Alzheimer's disease and gut microbiota	PubMed	Closed	Outlier detection, Cross-domain exploration	OntoGen, CrossBee	Manual review by experts is required	Development of a tool to provide recommendations for hypothesis generation, Semi-automated generation of ontologies, Use of term extraction
Kostoff et al. [36]	Identification of possible treatments for Inflammatory Bowel Disease	Medline	Open	Query formulation using biomarkers and theory desired treatment-derived directions of change	LRDI (Literaure Related Discovery and Innovation)	N/A	N/A
Chen et al. [37]	Detection of associations among complex diseases	PubMed abstracts	Open	Latent Semantic Analysis, Spectral clustering algorithm	SemMedDB (SemRep)	Performance could be improved by the use of deep learning	Use of large amount of training/testing data
Rindflesch et al. [38]	A plausible explanation for the correlation between epilepsy and inflammatory bowel disease	Medline titles and abstracts	Closed	Discovery Browsing	SemRep	SemRep is not accurate and has low values for precision and recall	Requirement to not only rely on semantic predications and manual inspection of citations
Dai et al. [39]	Identify candidate genes for the interaction between myocardial infarction and depression	Medline	Closed	ABC model of LBD	BITOLA	N/A	N/A
Rather et al. [40]	Discovery of potential new biomedical knowledge (relationships)	PubMed Abstracts, Clinical Trial protocols, NIHR grants summary	Open	Deep learning	Word2vec	N/A	Using a larger text corpus to find more meaningful and strong patterns, Exploratory analysis methods to discover hidden patterns

N/A means Not Available

4.3.1 Co-occurrence-based models

The ABC model of LBD is a common relation extraction technique used by many authors [18–20, 26, 30, 31, 39]. The associations between the different concepts are usually deduced from semantic predications extracted from NLP tools, like SemRep and MetaMap, which have been the most preferred tools. If the output of the ABC method consists of a long list of C terms, then these are ranked based on specific criteria and the higher-ranked C terms are considered as plausible hypotheses. Co-occurrence-based metrics are often used for analyzing the strength of entity associations, and prioritization of C terms are often based on the total frequency of co-occurrence [32]. Furthermore, Gubiani et al. [35] proposed a method to identify outlier documents by making use of two tools, namely OntoGen for outlier document detection and CrossBee for cross domain exploration.

Table 4 shows the different biomedical concepts A, B and C which have been considered in the studies in this review.

4.3.2 Distributional models

While most LBD methods apply co-occurrence-based methods to assess the relatedness of biomedical concepts, distributional models are also widely used. These models build vector representations of concepts which are based on the context in which they appear in literature. Relatedness between a pair of concepts is then derived based on the similarity between the vectors. Various distributional semantic techniques which have been proposed include Semantic Predications [18, 25, 30], Latent Semantic Analysis (LSA) [37], Predication-based Semantic Indexing (PSI) [28] and composite feature vectors [29]. Mower et al. [29] have shown that distributional models perform better than co-occurrence-based models.

4.3.3 Machine Learning models

Several authors have used machine learning in different steps of their LBD methodology. For text analysis, Pyysalo et al. [32] propose the use of machine learning-based methods for the recognition of biomedical entity names and their grounding to domain-specific ontology identifiers. Ranking of LBD-generated hypotheses have been performed by Zhang et al. [27] through a machine learning-based filter (lasso regression filter) and Rastegar-Mojarad et al. [21] by using a binary classifier. Machine learning algorithms like logistic regression [22, 23, 29] and k-Nearest Neighbor (kNN) [29] have been incorporated in models proposed by authors in this review. Rather et al. [40] integrated Word2vec, a neural network based algorithm, in their LBD approach and showed that the model was able to retrieve strong relationships which were not identified by UMLS. Deep learning has also been used in LBD techniques[18, 35].

Table 3 Main application areas for LBD in healthcare

Application Area	Number of studies
Drug repurposing	8
Pharmacovigilance and drug interactions	5
Identification of potential causes, therapies or treatments for specific diseases	6
Explanation for the correlation between diseases	3
Discovery of new biomedical knowledge (relationships)	1

4.3.4 Knowledge-graph models

Knowledge-graph models use graph theory to identify novel associations among various concepts. In their LBD approach for drug discovery, Zhao et al. [22] constructed a biomedical knowledge graph based on semantic predications. A path ranking algorithm was then used to extract drug-disease relation path features. Sang et al. [23] also use a knowledge graph-based drug discovery method, which involves the training of a logistic regression model by learning the semantic types of paths in the knowledge graph. Knowledge graph embedding and knowledge graph completion have also been used [24, 25].

4.4 RQ4: How do literature based discovery systems in the healthcare domain perform in terms of accuracy and reliability?

The papers analyzed have shown that diverse performance evaluation methods have been used for LBD systems, mostly due to the peculiarities of the healthcare domain and the specific requirements of the varied application areas.

4.4.1 No gold standard to benchmark performance

The evaluation of LBD systems in terms of accuracy and reliability is quite challenging in the healthcare domain. It becomes difficult for researchers to reliably distinguish between false positive signals and new discoveries. Most authors therefore have to rely on manual review by experts to confirm the final candidates for LBD. Many authors have claimed that there was no gold standard against which they could accurately benchmark the performance of their approaches [18, 21] and that precision and recall were not good metrics to measure the performance in all conditions [20].

4.4.2 Accuracy and reliability impacted by performance of text mining tools

The performance of the systems developed in several studies of this review is highly impacted by the performance of the tools and resources used in the LBD approach. In the evaluation of their system, Rastegar-Mijarad et al. [18] used the Comparative Toxicogenomics Database (CTD) resource, which does not annotate the type of relationship between drug and disease, therefore resulting in loss of valuable information. Sources of error are also often introduced in text mining tools like SemRep, due to inaccuracies in language processing or in the literature itself [21, 22, 25, 27, 38] and MetaMap whose accuracy reduces in the presence of ambiguity, resulting in the inability to resolve word sense disambiguation [23]. Sosa et al. [24] have acknowledged that the performance of their algorithm could considerably be improved if NLP tools improved their capability to capture complex relationships from unstructured text.

4.4.3 Computationally intensive models

The resource requirements for most LBD systems, specially those which use the open discovery approach are huge. Therefore, it is quite challenging for researchers to make their model computationally feasible, thereby imposing certain limitations resulting in suboptimal outcomes [24, 25, 28, 32]. One limitation of Pyysalo et al.'s [32] open discovery method is that it can recognize only a single correct target response for each case and their system is currently limited to discovery over paths of length two. Since the graph generated by the relations in SemMedDB is very large, making models computationally intensive, Zhang et al. [25] have used a sub-graph instead which resulted in loss of information, therefore affecting the accuracy of their model.

4.4.4 Limited data sets

Many authors agree that the use of larger and more variate data sets would improve the accuracy of their models. Limitations encountered include the use of unbalanced [21] and small [22, 32] data sets. The models proposed by Zhao et al.[22] and Pyysalo et al. [32] perform well using a rather small data set. However, the authors agree that their system's computational efficiency may be greatly reduced if the knowledge base is large. Yang et al. [19] believe that the rankings of the drug-disease pairs generated by their model may be adversely affected since their methodology did not consider aliases for drug names.

4.5 Research challenges and future directions

The proposed LBD approaches have demonstrated considerable achievements and promising results in the discovery process. An in-depth analysis of the techniques used has revealed major insights to the main research challenges and future directions for such systems. The proper handling of the research challenges will definitely result in improved accuracy and performance in the LBD process.

4.5.1 Minimize manual expert review

From the analysis of the various studies, it was found that extensive manual expert review was required for the selection of the final LBD candidates from a very large number. There is therefore the need to develop approaches to prioritize LBD candidates, which will provide domain experts with essential evidence instead of information overload. The following approaches are proposed to decrease the effort required by domain experts:

- Determine a suitable threshold score for LBD candidates [18, 21]. Candidates below that threshold would be considered as false positives and those above the threshold would be considered for further investigations and experiments.
- Develop a tool to provide recommendations for hypothesis generation [35]
- Make use of rigorous statistical techniques to replace the manual review step by a more automated approach [18]
- Design NLP techniques to detect false predications which occur due to negative associations [19–21, 23]

4.5.2 Seamless integration of multiple data sources for improved accuracy

Most models designed have only considered PubMED and MEDLINE abstracts as their main text corpus. Many authors have proposed the incorporation of additional data sources as the text corpus of their models to improve accuracy. A larger knowledge base has the potential to produce more complex relation paths. The additional data sources which could be considered include:

- NIH grants summary to identify potentially hidden and novel associations by investigating exploratory analysis methods [40]
- Biological data to find more drug candidates for Covid-19 drug repurposing [25]

Table 4 Biomedical concepts A, B and C considered in the ABC model of LBD

Study	Concept A	Concept B	Concept C	Type of Discovery	Discovery
Meng et al. [31]	Stroke	Assessment Scales	Rehabilitation Therapy	Open	Hand-arm bimanual intensive training (HABIT) was found to be a promising rehabilitation therapy for stroke
Rastegar-Mojarad et al. [18]	Drug	Gene	Disease	Open	Potential novel drug-disease pairs
Rindflesch et al. [38]	Inflammatory Bowel Disease (IBD)	Interleukin-1 beta and glutamate	Epilepsy	Closed	Interleukin-1 beta influence on glutamate levels is involved in the etiology of both IBD and Epilepsy
Yang et al.[19]	Disease	Gene	Drug	Open	Potential anticancer drugs
Xie et al. [26]	Drug	Indication (depression) / Side Effect	Herb (Traditional Chinese Medicine)	Open	The herb Pogostemon Cablin Benth can be an alternative to the drug Nefazodone, since it can mitigate the side effects
Raja et al. [20]	Disease	Phenotypes, symptoms	Drug	Open	Potential drugs identified for four diseases
Pyysalo et al. [32]	Arsenic	Nrf2 Gene	Autotaxin Protein	Closed	The properties of the Nrf2 gene explained the connection between arsenic and the autotaxin protein
Zhang et al. [27]	Cancer drug	Gene	Dietary supplement	Closed	Echninacea was found to be the first drug supplement interaction candidate area of interest
Gubiani et al. [35]	Alzheimer's disease	Chemicals, mechanisms of action, cell components	Gut microbiota	Closed	Nitric Oxide Synthase was found to be a promising novel bridging term for the neuronal and immunity field
Rastegar-Mojarad et al. [21]	Drug	Gene	Disease	Open	Potential novel drug-disease pairs
Sang et al. [23]	Disease	Protein	Drug	Open	Potential novel disease-drug pairs
Dai et al. [39]	Myocardial Infarction (MI)	Gene, gene product	Depressive disorder	Closed	Genes GNB3, CNR1, MTHFR and NCAM1 were found to be new putative candidate genes that may influence the interactions between MI and depression
Hristovski et al. [30]	Drug	Gene, protein	Adverse effect	Closed	Explanation for the association between drug and adverse effect through linking genes or proteins
Zhang et al. [25]	Drug	Any concept	Disease (Covid-19)	Open	A list of potential drugs for Covid-19

- Biomedical ontologies to consider additional interesting associations [38]
- Drug-disease databases like CTD and DrugBank for better training in drug-repurposing [19]
- FAERS data for pharmacovigilance methods instead of only relying on EHR data [28]
- Spontaneous reporting data for the extraction of drug-side effect associations [29]

4.5.3 Computational optimisation for improved accuracy and reliability

Studies in this review have clearly indicated the quest for researchers to obtain more accurate results. Due to the very large datasets and the multitude of possible pathways, the LBD models proposed are computationally intensive, therefore leading to certain limitations. Techniques proposed to improve accuracy include:

- Integration of machine learning and deep learning algorithms in LBD models [27, 29, 37]
- Development of high-quality NLP tools for better accuracy, due to the reported shortcomings of existing tools
- Use of relevant tools for the normalization of gene and disease targets [19]
- Consideration of full texts of research articles instead of only titles and abstracts [32]
- Use of graph embedding to obtain long paths [23]
- Consideration of indirect relationships from knowledge graphs [24]

5 Conclusion

The purpose of this work was to carry out a systematic literature review of recent publications in Literature Based Discovery approaches in the field of evidence-based healthcare. Four research questions had been set out in the planning phase of the review and the papers were deeply analyzed so as to get insights on the research questions. This work has revealed the potential of LBD techniques to discover hidden knowledge in emerging areas of healthcare and provides a comprehensive contextualization to various stakeholders in the health informatics community. The results of this review will therefore help the latter to have a good understanding of the appropriate approaches used in different application areas and contexts, and the challenges they will have to face.

The synthesis of the results of this work has shed light on recent issues and challenges that drive new LBD models and provides avenues for their application in other diverse areas in the healthcare domain. The research challenges identified show different perspectives to address further research in the field and, if properly tackled, will result in better overall accuracy and performance of LBD systems, therefore contributing in the speeding up of the knowledge translation process.

Authors' Contributions All authors have made a substantial, direct, intellectual contribution to this study.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability Not applicable.

Code availability Not applicable

Declarations

Conflicts of interest None.

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

References

1. Moses H, Matheson DH, Cairns-Smith S, George BP, Palisch C, Dorsey ER. The anatomy of medical research: US and international comparisons. Jama. 2015;313(2):174-89.
2. Graham ID, Tetroe J. How to translate health research knowledge into effective healthcare action. Healthc Q. 2007;10(3):20–2.
3. Blair M. Getting evidence into practice—implementation science for paediatricians. Arch Dis Child. 2014;99(4):307–9.
4. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med. 2011;104(12):510–20.
5. Shayan SJ, Kiwanuka F, Nakaye Z. Barriers associated with evidence-based practice among nurses in low-and middle-income countries: A systematic review. Worldviews on Evidence-Based Nursing. 2019;16(1):12–20.
6. Sebastian Y, Siew EG, Orimaye SO. Emerging approaches in literature-based discovery: techniques and performance review. The Knowledge Engineering Review. 2017;32.
7. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering (2007).
8. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–9.
9. Pearson A, Jordan Z, Munn Z. Translational science and evidence-based healthcare: a clarification and reconceptualization of how knowledge is generated and used in healthcare. Nurs Res Prac. 2012.
10. Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. 1996.
11. Preiss J, Stevenson M, Gaizauskas R. Exploring relation types for literature-based discovery. J Am Med Inform Assoc. 2015;22(5):987–92.
12. Henry S, McInnes BT. Literature based discovery: models, methods, and trends. J Biomed Inform. 2017;74:20–32.
13. Hristovski D, Rindflesch T, Peterlin B. Using literature-based discovery to identify novel therapeutic approaches. Cardiovascular & Hematological Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Cardiovascular & Hematological Agents). 2013;11(1):14-24.
14. Kim YH, Song M. A context-based ABC model for literature-based discovery. PloS One. 2019;14(4):e0215313.
15. Lever J, Gakkhar S, Gottlieb M, Rashnavadi T, Lin S, Siu C, Smith M, Jones MR, Krzywinski M, Jones SJ. A collaborative filtering-based approach to biomedical knowledge discovery. Bioinformatics. 2018;34(4):652–9.
16. Meier JJ, Conkling TW. Google Scholar's coverage of the engineering literature: an empirical study. J Acad Librariansh. 2008;34(3):196–201.

17. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics. 2019;118(1):177–214.
18. Rastegar-Mojarad M, Elayavilli RK, Li D, Prasad R, Liu H. A new method for prioritizing drug repositioning candidates extracted by literature-based discovery. In2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2015;69-674. IEEE.
19. Yang HT, Ju JH, Wong YT, Shmulevich I, Chiang JH. Literature-based discovery of new candidates for drug repurposing. Brief Bioinform. 2017;18(3):488–97.
20. Raja K, Steill J, Ross I, Tsoi LC, Kuusisto F, Ni Z, Livny M, Thomson J, Stewart R. SKiM-A generalized literature-based discovery system for uncovering novel biomedical knowledge from PubMed. bioRxiv. 2020.
21. Rastegar-Mojarad M, Elayavilli RK, Wang L, Prasad R, Liu H. Prioritizing adverse drug reaction and drug repositioning candidates generated by literature-based discovery. InProceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics 2016;289-296.
22. Zhao D, Wang J, Sang S, Lin H, Wen J, Yang C. Relation path feature embedding based convolutional neural network method for drug discovery. BMC Med Inform Decis Mak. 2019;19(2):59.
23. Sang S, Yang Z, Wang L, Liu X, Lin H, Wang J. SemaTyP: a knowledge graph based literature mining method for drug discovery. BMC Bioinformatics. 2018;19(1):193.
24. Sosa DN, Derry A, Guo M, Wei E, Brinton C, Altman RB. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. InPac Symp Biocomput. 2020;25:463–74.
25. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. J Biomed Inform. 2021;115:103696.
26. Xie Q, Yang KM, Heo GE, Song M. Literature based discovery of alternative TCM medicine for adverse reactions to depression drugs. BMC Bioinformatics. 2020;21(5):1–9.
27. Zhang R, Adam TJ, Simon G, Cairelli MJ, Rindflesch T, Pakhomov S, Melton GB. Mining biomedical literature to explore interactions between cancer drugs and dietary supplements. AMIA Summits on Translational Science Proceedings. 2015;69.
28. Malec S, Gottlieb A, Bernstam E, Cohen T. Using the literature to construct causal models for pharmacovigilance. Easy Chair. 2018;23.
29. Mower J, Subramanian D, Cohen T. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. J Am Med Inform Assoc. 2018;25(10):1339–50.
30. Hristovski D, Kastrin A, Dinevski D, Burgun A, Žiberna L, Rindflesch TC. Using literature-based discovery to explain adverse drug effects. J Med Syst. 2016;40(8):185.
31. Meng G, Huang Y, Yu Q, Ding Y, Wild D, Zhao Y, Liu X, Song M. Adopting literature-based discovery on rehabilitation therapy repositioning for stroke. BioRxiv. 2018:422154.
32. Pyysalo S, Baker S, Ali I, Haselwimmer S, Shah T, Young A, Guo Y, Högberg J, Stenius U, Narita M, Korhonen A. LION LBD: a literature-based discovery system for cancer biology. Bioinformatics. 2019;35(9):1553–61.
33. Kostoff RN, Patel U. Literature-related discovery and innovation: chronic kidney disease. Technol Forecast Soc Chang. 2015;91:341–51.
34. Gubiani D, Petrič I, Fabbretti E, Urbančič T. Mining scientific literature about ageing to support better understanding and treatment of degenerative diseases. In Conference on Data Mining and Data Warehouses. Ljubljana 2015.
35. Gubiani D, Fabbretti E, Cestnik B, Lavrač N, Urbančič T. Outlier based literature exploration for cross-domain linking of Alzheimer's disease and gut microbiota. Expert Syst Appl. 2017;85:386–96.
36. Kostoff RN, Briggs MB, Shores DR. Treatment repurposing for inflammatory bowel disease using literature-related discovery and innovation. World J Gastroenterol. 2020;26(33):4889.
37. Chen G, Jia Y, Zhu L, Li P, Zhang L, Tao C, Zheng WJ. Gene fingerprint model for literature based detection of the associations among complex diseases: a case study of COPD. BMC Med Inform Decis Mak. 2019;19(1):1–9.
38. Rindflesch TC, Blake CL, Cairelli MJ, Fiszman M, Zeiss CJ, Kilicoglu H. Investigating the role of interleukin-1 beta and glutamate in inflammatory bowel disease and epilepsy using discovery browsing. Journal of biomedical semantics. 2018;9(1):25.
39. Dai Z, Li Q, Yang G, Wang Y, Liu Y, Zheng Z, Tu Y, Yang S, Yu B. Using literature-based discovery to identify candidate genes for the interaction between myocardial infarction and depression. BMC Med Genet. 2019;20(1):104.
40. Rather NN, Patel CO, Khan SA. Using deep learning towards biomedical knowledge discovery. Int J Math Sci Comput (IJMSC). 2017;3(2):1–10.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.