


REVIEW

Automated data analysis of unstructured grey literature in health research: A mapping review

Lena Schmidt¹  | Saleh Mohamed¹ | Nick Meader¹ | Jaume Bacardit² | Dawn Craig¹

¹National Institute for Health and Care Research Innovation Observatory, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK

²Interdisciplinary Computing and Complex BioSystems (ICOS) Research Group, School of Computing, Newcastle University, Newcastle upon Tyne, UK

Correspondence

Lena Schmidt, National Institute for Health and Care Research Innovation Observatory, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK.

Email: lena.schmidt@io.nihr.ac.uk

Funding information

National Institute for Health and Care Research, Grant/Award Number: HSRIC-2016-10009

Abstract

The amount of grey literature and ‘softer’ intelligence from social media or websites is vast. Given the long lead-times of producing high-quality peer-reviewed health information, this is causing a demand for new ways to provide prompt input for secondary research. To our knowledge, this is the first review of automated data extraction methods or tools for health-related grey literature and soft data, with a focus on (semi)automating horizon scans, health technology assessments (HTA), evidence maps, or other literature reviews. We searched six databases to cover both health- and computer-science literature. After deduplication, 10% of the search results were screened by two reviewers, the remainder was single-screened up to an estimated 95% sensitivity; screening was stopped early after screening an additional 1000 results with no new includes. All full texts were retrieved, screened, and extracted by a single reviewer and 10% were checked in duplicate. We included 84 papers covering automation for health-related social media, internet fora, news, patents, government agencies and charities, or trial registers. From each paper, we extracted data about important functionalities for users of the tool or method; information about the level of support and reliability; and about practical challenges and research gaps. Poor availability of code, data, and usable tools leads to low transparency regarding performance and duplication of work. Financial implications, scalability, integration into downstream workflows, and meaningful evaluations should be carefully planned before starting to develop a tool, given the vast amounts of data and opportunities those tools offer to expedite research.

KEYWORDS

artificial intelligence, automation, grey literature, literature review, natural language processing

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

Highlights

What is already known

- There is a time lag between novel developments of technologies, versus their publication in peer-reviewed literature and finally their appearance in systematic reviews.
- The inclusion of grey literature can help to overcome this time lag, but the amount of data can be overwhelming or not straightforward to access.

What is new

- Automation through Natural Language Processing (NLP) can enable the analysis of grey literature at scale.
- A total of 84 papers for 7 tools and 76 methods were included in this review, covering different sources of grey literature and the most common data points for extraction or analysis.

Potential impact for Research Synthesis Methods readers

- Readers with an interest in developing automation methods will gain an overview of the state of NLP research and datasets.
- Readers with an interest in using automation methods will gain an overview of tools, their features and performance evaluations.

1 | INTRODUCTION

1.1 | Background

The literature landscape in health and social care is evolving rapidly. Research outputs are being published at an unprecedented rate, which in turn has increased the rate and scale of secondary research projects, such as systematic reviews, rapid reviews, evidence gap maps, and horizon/future pipeline scans. Published and peer-reviewed literature, among other types of data, can provide important evidence used to inform choice and implementation of medicines or medical devices within a healthcare system.

However, there is a time lag between novel developments of technologies and associated research, versus their publication in peer-reviewed literature. Published research often become available years after the development of a medicine or technology. Analyses estimate the peer-review to publication time-lag for medicine and medical device trials alone as up to 4–7 years^{1–3}; and fully relying on systematic review processes to support decision making would lead to further delays. In a recent analysis of 20,000 systematic reviews, DeYoung et al.⁴ found that the median delay between study and review publication was another additional 8 years,⁴ which explains why other types of secondary literature review, such as HTA or horizon scans also use non-peer-reviewed information, to create better

representations of current developments and their early evidence base.^{5,6}

To enable a comprehensive analysis of current and ongoing developments, there is a growing need to explore and consider these ‘softer’ sources of intelligence, often using a combination of grey literature and other health-related information in the public domain. Grey literature itself can be defined as ‘that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers’.^{7,8} It includes information from sources such as clinical trial registries, preprint servers, or academic outputs such as conference proceedings, dissertations, and theses. These types of grey literature can be very valuable to health-related secondary research by uncovering first traces of new or ongoing research or collecting additional information about studies already found in peer-reviewed literature.^{8,9}

However, this can also include industry-focused or legal texts such as patents or websites, or reports from governments and charities. News articles and press-releases are yet another example for public-domain and non-peer-reviewed sources of health-related information that can be considered as softer grey literature⁷; alongside social media sites, which have the potential to provide intelligence closest to real-time development of innovations in healthcare. These latter soft data sources have not been traditionally counted as grey literature.

Figure 1 seeks to illustrate the process between very early-stage research and adoption into practice on a

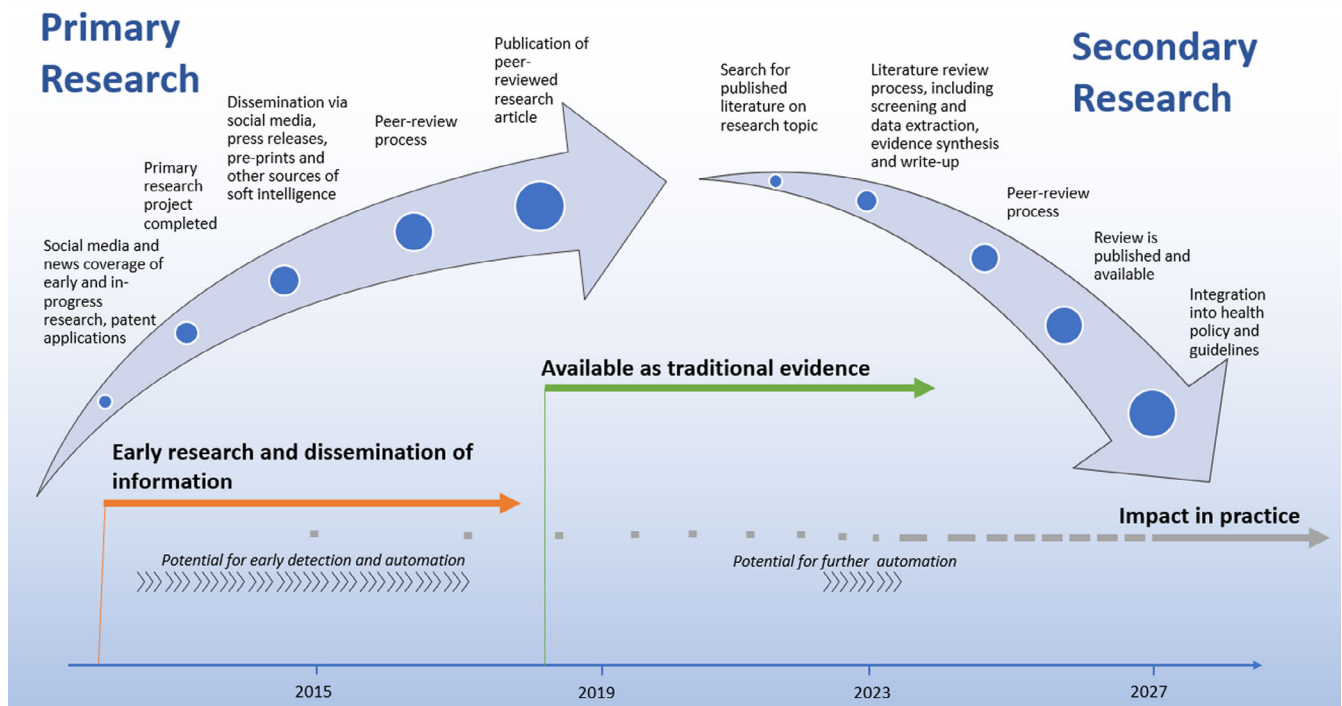


FIGURE 1 Exemplary timeline of the development and adoption of research into practice. [Colour figure can be viewed at wileyonlinelibrary.com]

very high level. It shows where, in theory, the scope of secondary research could be extended to include novel sources of information for an earlier detection of potentially relevant research trends. It also shows areas where automation in earlier retrieval of evidence could potentially help to accelerate discovery of information, and where automated data extraction could help to make the process of uncovering intelligence from soft data more efficient.

Research areas beyond the scope of classic systematic reviews, for example horizon scanning activities or HTA, utilise traditional sources of evidence (i.e., clinical trials or diagnostic accuracy studies), but also benefit from including novel sources of information from the public domain to detect signals of future trends and maturing technologies in a timely manner.^{5,6} In the scope of a systematic review, Hines et al⁶ mapped data sources used in horizon scanning and found that softer sources of intelligence used in published projects included for example patents, surveys, or media content to detect likely technological trends in the near future.

However, there exist challenges in using such data. Those challenges go back to fundamental differences between traditional, peer-reviewed evidence-based information and softer information about early research developments. The differences can be negative, in terms of lower quality of the information

content.¹⁰ But they can also be positive, in terms of rapid dissemination and availability of data that would usually be held-up in lengthy clinical trials or peer-review processes.

It is critical that methods of information retrieval and data extraction advance to keep pace with this infodemic. This is especially true when expanding the scope of data sources that are used for secondary literature analysis into the domain of grey literature and soft data. These information retrieval methods underpin the screening process for relevant literature and data extraction in research. Automation has a key role to play in providing faster and more resource-efficient evidence synthesis whenever the impact on general health or the implication of a medicine, therapy, or technology within a healthcare system overall is unclear.

In their 2021 survey of HTAs, which include grey literature,⁵ the WHO noted that >70% of the 127 included countries used HTAs to plan, budget, and to inform clinical practice guidelines.¹¹ Stakeholders involved in the process of prioritising and creating these HTAs are government entities and national health services, as well as patient organisations or industry. However, two of the main barriers to the production of assessments such as HTAs are budget and data availability,¹¹ representing bottlenecks that can be addressed via the usage of natural language processing (NLP) and automated information retrieval and extraction.^{6,12}

TABLE 1 Examples for types of grey literature information and exemplary data sources.

Information type	Examples for data sources
Social media	Twitter, Reddit, YouTube
Internet fora	Mumsnet, SANE
News	Google News, Med-Tech News, PharmaTimes
Government Agencies or charities	Websites, eg UNICEF, world bank
Patents	Databases and indexes, such as USP
Clinical trial registries	ClinicalTrials.gov and other registries
Pre-prints	MedRxiv, ArXiv

1.2 | Aims

This paper provides an overview of automated data extraction methods and tools for health-related research questions that can be answered using grey and soft data. We discuss the sources from which data are automatically extracted (e.g., social media, patents, news) and the type of data that are extracted (e.g., diseases, drugs, technologies). Among other items we cover performance, practical value, as well as challenges and barriers to the implementation of automated data extraction methods.

1.3 | Related research

With advances in NLP and developments in deep- and machine-learning, it is becoming feasible to process vast amounts of unstructured digitalized texts. This is giving rise to the emerging field of NLP-based health data science, where novel research in data mining and data extraction is specifically applied to automate work in evidence synthesis. A living systematic review of automated data extraction from the highly related field of peer-reviewed health literature currently includes 76 papers, indicating fast-paced advancements in the areas of automatic extraction, normalisation, relation extraction, and text summarisation.¹³ Within these advancements, there remains a need to explore methods of automatically processing unstructured text data in the non-peer-reviewed space, and to assess which tools and methods will facilitate this process for end-users. NLP and text mining is frequently used to analyse or extract data from social media platforms such as Twitter.¹⁴ Applications range from vehicle traffic analysis¹⁴ to medical and health data.¹⁵ Correia et al¹⁶ published a narrative review of

TABLE 2 Overview of databases.

Database	Interface or URL
MEDLINE via PubMed	https://pubmed.ncbi.nlm.nih.gov
Scopus	https://www.scopus.com
ACL Anthology	https://aclanthology.org/
dblp: computer science bibliography	https://dblp.org/
MedRxiv	https://www.medrxiv.org/
ArXiv (computer science)	https://arxiv.org/archive/cs

recent work on data mining in social media content analysis. They discuss papers on automation in the domains of pharmacovigilance and sentiment analysis, most commonly targeting specific drugs and their adverse events, or mental health research questions. A large amount of related research has been conducted on information extraction from electronic health records, for example extracting diagnoses, treatments,¹⁷ or genomic data.¹⁸ Other grey literature data sources such as pre-prints¹⁹ and clinical trial registrations are targets for data mining and extraction to connect them with their published counterparts.^{20,21}

2 | METHODOLOGY

2.1 | Research objective

This review maps published tools and methods for literature mining and data extraction. A ‘tool’, in this context is defined as an end-user application with a user-interface, available for example as web or desktop application. A ‘method’ is defined as a set of scripts or a description of an algorithm that requires users to be familiar with data science or programming. Results of this literature review were summarised in the form of an evidence map, visualising the extracted data, current knowledge, and research gaps. The review includes any publications that describe approaches to expedite data extraction from grey literature and soft data. In this review, grey literature and soft data includes any health-related data that has not passed peer-review; with examples given in Table 1.

Considering open questions around usefulness, feasibility, and practical integration of grey literature and soft data, the motivation for this literature review is to identify and examine tools and methods that currently exist and have been used to automate data extraction activities from these publicly available data sources.

2.2 | Literature searches

A robust search strategy was developed to identify relevant articles from a variety of electronic databases, covering health, informatics, and pre-prints in both health research and informatics. The initial search strategy was developed using the PubMed 'Advanced Search' function. Six databases were searched (2005–2022), each using a database-specific adaptation of the PubMed search: MEDLINE (searched via PubMed); Scopus; ACL; dblp computer science bibliography; MedRxiv; and ArXiv (see Table 2).

The start date of 2005 was selected, since this is the year after which publications relevant to text-mining in general systematic review automation first started to appear. Three published systematic reviews of data extraction methods from the related field of peer-reviewed literature did not find any published text-mining or data extraction approaches prior to 2005.^{13,22,23} We furthermore decided to keep this 2005 date filter because the availability of data sources changes over the years, and methods published prior to this date are not representative anymore or are becoming unlikely to be usable in practice due to changes and updates in programming languages (i.e., new Python¹ or Java² versions).

The PubMed search strategy was developed, and refined further based on feedback from an independent information specialist. The strategy was then adapted for usage in Scopus. Searches on the ACL, dblp, MedRxiv, and ArXiv were adapted and carried out as described by McGuinness and Schmidt.²⁴ In short, we utilised full database exports of all papers indexed by these databases, and then used methods from the medrxivr R package³ to retrieve relevant records. Search strategies, including the regular-expression-based search for the ACL and pre-print servers, are included as Appendix D in Data S1 (see online supporting information).

TABLE 3 Screening criteria.

Inclusion criteria: title and abstracts	General exclusion criteria: publications with datasets focussing on
Describes original data extraction tool or method	Patient level data such as electronic health records
Uses at least one dataset with non-peer-reviewed data related to healthcare	Genomic or biological data extraction such as gene expressions or proteins
Inclusion criteria: full texts	
Published full texts, such as journal, conference, or pre-print papers	
Publication available in English	

2.3 | Eligibility criteria

In addition, to these inclusion criteria described in Table 3, during title and abstract screening we separately tagged papers that describe the usage or evaluation of a tool or method with respect to a specific health research question. For this, we tagged two items:

1. The topic of research: We created a vocabulary to categorise and bin the specific health topic studied in the reference, based on information available in the title and abstract. The tags included, for example, mental health or Covid-19.
2. The data sources: We created a vocabulary to categorise and bin the sources of mined data. The tags included, for example, Twitter or health-related fora.

The decision to tag, but then exclude topic-specific research papers at the title/abstract level was made after a pilot-study showed that full inclusion of every such paper would lead to an unfeasibly large amount of included papers. We imported all tags into the SWIFT-Review software,²⁵ to create visualisations in the form of heatmaps, bar- and pie-charts and to make the whole dataset publicly sharable. A description of these results is given in Appendix C in Data S1.

2.4 | Screening and workflow management

All papers were deduplicated, screened, and data-extracted in SWIFT-ActiveScreeener.²⁶ Screening at title/abstract level was conducted up to an estimated sensitivity of 95% by one reviewer, and a second reviewer independently checked random samples of in- and excluded records. Howard et al²⁶ describe this in more detail, and additionally we added an in-depth explanation of this process in the [Supporting Information](#). Conflicts were discussed and resolved until the

TABLE 4 Search results.

Database	Number of results
PubMed	2108
MedRxiv	89
dblp	81
ArXiv	1332
ACL	387
Scopus	5704

screeners were confident that the in- and exclusion criteria were applied correctly.

Similar to the initial screening process, full-text screening and data extraction decisions were reviewed by an independent reviewer. Conflicts were discussed and resolved in the same fashion.

Where there were multiple publications describing the development or evaluation of the same tool, we grouped those papers and jointly extracted data once for each tool, focusing on the most recent version of each feature or function.

2.5 | Data extraction

Data were extracted within SWIFT-ActiveScreeener. The data extraction questionnaire was set up in the form of

text fields and checkboxes, as applicable. In the following we provide an overview of the extracted data, the full questions are shown in Appendix A in Data S1.

- We extracted relevant tool features and functionalities from each paper, a list of data sources from which information was obtained (e.g., GoogleNews or Twitter), and the type of data (e.g., patents or online fora).
- We extracted whether the paper refers to a tool or method, the extent of automation of the analysis (e.g., recognition of entities or full normalisation to standardised vocabularies), metrics and methods used for validation within the paper (e.g., F1, precision), and description of the tool's integration into real research projects, where applicable.
- Finally, we extracted any challenges or barriers related to the development or deployment of tools and

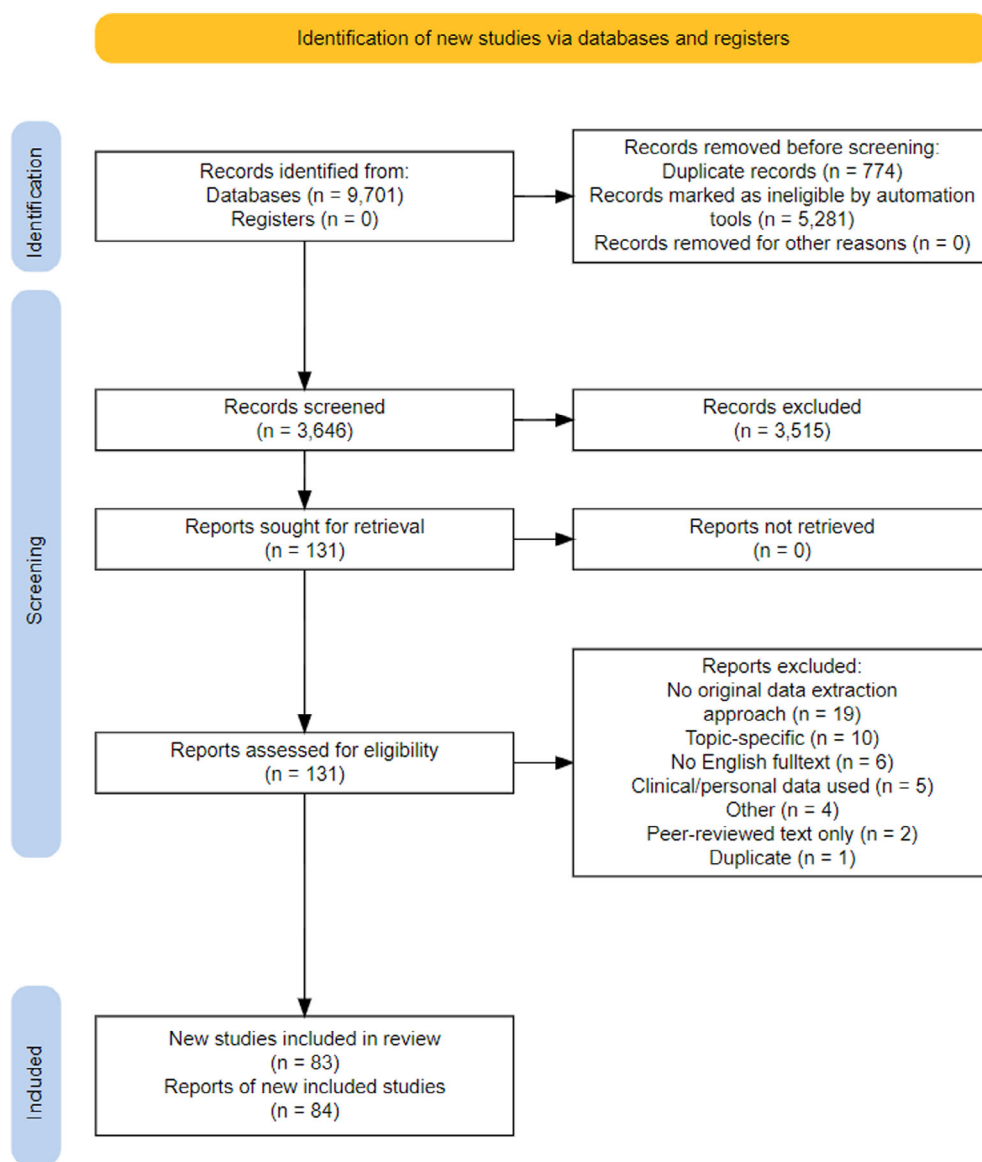


FIGURE 2 PRISMA2020 flow diagram.^{27,28} [Colour figure can be viewed at wileyonlinelibrary.com]

methods, caveats when using these in real-world research projects, and research gaps described in the papers.

3 | RESULTS

3.1 | Screening

In total, the searches retrieved 9701 references (databases searched up to 03/2022, see Table 4). After deduplication, 8927 references were imported into SWIFT-ActiveScreeener for screening.

As described within the method section of this paper, screening was conducted using early stopping at a target estimated sensitivity of 95%.²⁶ After reaching the target sensitivity a further 1000 references were screened, but screening was then stopped because no further relevant records were identified.

In total, 3646 titles and abstracts were screened. Of those, 318 titles and abstracts were excluded at title and abstract level, but still tagged by research topic and data source because they described topic-specific analyses conducted based on non-peer-reviewed data. These tags were only applied during the abstract screening process, and the 318 references did not proceed to full text screening.

On full text we included 83 tools and methods. Eighty-four papers were included, but two were grouped together because they described the same tool (see Figure 2).

3.2 | Summary of the full-text literature

We imported all data into the SWIFT-Review⁴ software,²⁵ to create visualisations in the form of heatmaps, bar- and pie-charts. The project file (.stp format) is shared in the [Supporting Information](#), to make the whole dataset publicly accessible for free.

A total of 84 papers for 7 tools and 76 methods were included at the full-text level. One tool was described by two papers. Data were extracted into extraction forms created within the screening tool.

In the following section, we firstly focus on results from tool papers or method papers describing tool design and deployment. In the section, thereafter, we focus on methods papers, evaluation of methods and more technical details of algorithms doing automated data summarisation, analysis, or normalisation. The final results section focuses on practical challenges and research gaps in deploying and using automation tools.

TABLE 5 End-user tools for automated data analysis.

Title	Tool name	Data	Deployment	References	Source	Link to tool
A user-friendly tool for medical-related patent retrieval	TWINC	Patents	Unclear	³¹	PubMed	
PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance	PADI-web	News	Web	³⁰	PubMed	https://padi-web.cirad.fr/en/
A new visual navigation system for exploring biomedical Open Educational Resource (OER) videos	-	Video	Unclear	³²	PubMed	
Mining Adverse Drug Reactions from Unstructured Mediums at Scale	-	Social Media, EHR	Unclear, some code given	³³	ArXiv	
Development and evaluation of a prototype search engine to meet public health information needs	PHIS	Websites	Unclear	³⁴	PubMed	
iPresage: An innovative patent landscaping tool	iPresage	Patents	Web	³⁵	Scopus	
E-patent examiner: Two-steps approach for patents prior-art retrieval	E-patent examiner	Patents	Web	³⁶	Scopus	

3.2.1 | Tool descriptions, features, and integration into review workflows

Tool descriptions

We found 7 end-user tools and 76 published methods papers. For one tool (PADI-web), we aggregated two full texts into one tool description.^{29,30} Those tools automate data extraction across different types of text (i.e., patents, news, trial registrations) and across different media (i.e., digitalised text and videos). For PADI-web, we found accessible web-deployments, giving users the opportunity to test and use the tool. Code has been published for one further tool's NLP models. The remaining tools were not accessible online and we were unable to find publicly accessible deployments or executable desktop applications.

Table 5 below shows an overview of the tools, key information about them and a short description.

Tool features

A total of 16 papers described features and functionalities implemented in tools, or features that could be of use in practice when processing the data via a method. Pasche et al³¹ noted the ability to bulk-process data and the utility of automatic query expansion, for example to automatically increase the amount of chemical terms by adding synonyms found within MeSH or Pubchem terminologies within their tool called 'TWINC'. As part of the papers for the PADI-web tool, Valentin et al³⁰ described the feature of automatic daily evidence updates to the data, via RSS feeds, crawling of related websites, and usage of the Google News API. The tool can retrieve new data and therefore prioritise, mine, and normalise new information as it is published; ensuring that research-projects are up-to-date. They support data annotation via an integration of the BRAT tool³⁷ and make the tool publicly available as web-application, thus facilitating collaboration and re-use of manually extracted data. They furthermore describe features such as email-notifications and summaries sent to the user, which helps with transparently communicating changes in the evidence and providing fast and easy-to-digest updates without accessing the tool itself every day. Hari-prasad et al³⁸ discuss added value of a user-interface to visualise automatically mined or extracted data, in the form of histograms, pie charts or other types of plots.

Natsiavas et al³⁹ describe the user-requirements and design-process of a future tool, citing the full pipeline of prioritisation/mining/normalisation as a feature. They discuss the problem of heterogeneity between data sources and suggest a division to explore data from different sources separately, as well as separate data mining and normalisation for each data source. As final, separate

feature, they describe a data consolidation process that includes automated reports and visualisations to follow-up on new data. In the PHIS tool, which stands for Public Health Information Search,³⁴ public-health websites are crawled and there can be a focus on more than one class of entities. Documents summaries are provided with respect to extracted data. Tafti et al⁴⁰ describe a data mining architecture for mining social media data, and note that usage of their database-infrastructure as a feature to increase scalability and future access via a tool.

Lee and Uzuner⁴¹ processed patent data and described benefits of a feature to divide patents between already-commercialised products and between technologies in development. Also in the patent-space, Avasarala and Bonissone³⁵ (iPresage tool) describe colour-coding patent-assignees for better identification and visualising temporal trends via stacked histograms. The E-patent examiner tool by Kravets et al³⁶ is also a patent-focused web-application, citing being web-based as a positive feature, as well as allowing expert-input on top of the automated process.

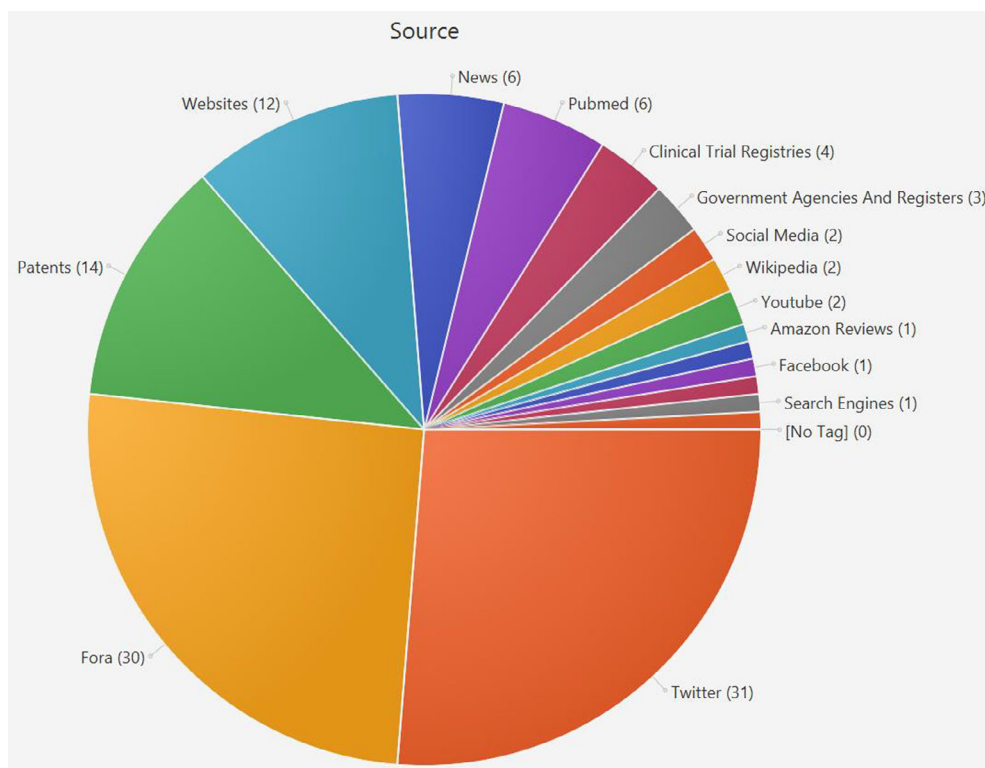
For video-data, Zhao et al³² (unnamed tool) describe implementing features that help users gain a streamlined overview of the data. This includes automatically indexing and updating their dataset with new health-related videos, similar to the updates within PADI-web. They also mention benefits of making the tool available as web-application. The video-specific features include visualising mined content as part of the video timeline and making it easy to skip between highly relevant sections, using hover text and visual cues such as word-clouds that represent key moments. Multiple videos can be visualised on the same grid for comparison. In terms of user-management, they discuss features to add comments to a video and user-account management.

Tool usage within review workflows

Four included papers described practical settings. Tasks in which a tool or method was used included information gathering and scoping before starting a review project, usage during the data extraction phase, or in the scope of clinical practice.

The PADI-web 3.0 tool³⁰ describes integration into practice both in terms of scoping and in terms of keeping researchers up-to-date automatically, on a daily basis. When describing the scoping process, they note that one way to integrate automation of grey literature data extraction is by using it to 'triage' information before further review. Triage itself is a term describing prioritisation of patients in emergency situations, to administer treatment first to anyone who might benefit from it most. Similarly, researchers in health-related topics focus on the best available, peer-reviewed evidence first and then

FIGURE 3 Sources of data used in the included papers. We only included full-texts if they reported usage of such data within their dataset. PubMed was tagged as data source whenever an included reference mentioned using a mixed corpus such as TwiMED, including a mix of PubMed and Twitter data. One paper might include more than one source of data. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]



a system of triage for less strong sources of evidence, such as news and social media, may be applied. With a focus on the field of disease surveillance, Valentin³⁰ note that classifying and removing likely irrelevant data from additional sources of information via machine-learning is a step that brings value to a research project because it leads to a prioritised and therefore earlier detection of information. The practical integration into review workflows happens via automated dissemination of content to the reviewers. Specifically, automated email-updates, summarisation of newly classified relevant information, and usage of RSS-feeds is described.³⁰

Similar to PADI-web 3.0, Zhao et al³² describe a tool for information scoping that is not directly connected to feed information into review tools. Using biomedical educational videos, they support the process of scoping by helping to apply information-specialist curated keyword searches and skimming video content. In practice, this aims to prioritise the display of likely relevant content and thus reducing time needed to watch the full content. However, similar to PADI-web, no export of the classified or extracted information into downstream evidence-synthesis tools is described.

Turner et al⁴² developed a data model describing key pieces of evidence extracted from grey public-health data, for downstream usage in automation tools. They characterised information needed in practice and discuss using a rule-based approach for automatic information extraction that fits their data model.

In contrast to the common focus on literature reviews, Natsiavas et al³⁹ discuss requirements for the integration of new information into a clinician's workflow. They do not describe integrating with downstream tools, but describe value added for clinicians by providing automated summaries and analyses of texts describing adverse events. They mine data, normalise and consolidate, provide structured reports and follow up on new data extracted from social media, government websites such as FDA, and patient health records.

3.2.2 | Analysis of automation methods: Data sources, types of data, extend of automation, and evaluation

Data sources

We tagged the source of data used within the included full-texts and show the results in Figure 3, using the same set of tagging categories that was also applied to the topic-specific datasets analysed on title and abstract-level in Appendix C in Data S1. The results within Figures A1 and A2 (topic-specific abstracts, see Appendix C in Data S1) and Figure 3 (included full-texts) are very similar, both indicating that Twitter, health-related fora, websites and news are the most common sources of data used for automation.

For the purpose of training and evaluating an algorithm, researchers often use publicly available

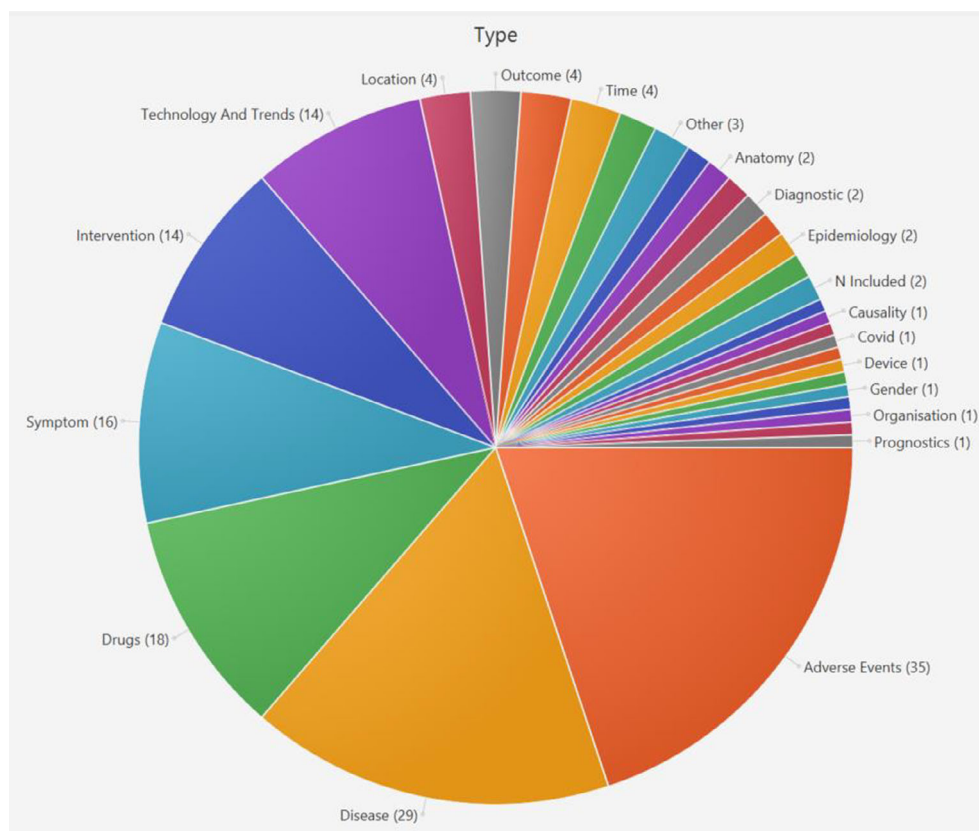


FIGURE 4 Types of data covered by automation methods in the included tools and papers. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1692)]

TABLE 6 Level of automation in the included papers. One paper can include more than one type of automation.

	2005– 2006	2007– 2008	2009– 2010	2011– 2012	2013– 2014	2015– 2016	2017– 2018	2019– 2020	2021– 2022
Priorisation and Summarisation	1	0	2	5	3	8	15	16	8
Mining Entities and Sentences	0	0	0	1	3	10	8	15	8
Extraction and Normalisation	0	0	0	0	1	2	6	5	3

benchmarking corpora or they create custom labelled datasets for this purpose.

We found 25 different benchmark corpora used as data sources for training and/or evaluation in the included papers. The most commonly used corpora were SMM4H⁵ used by 8 papers, and CADEC⁴³ and TwiMED,⁴⁴ used by 5 papers each. One corpus was used by two papers, and the remaining 21 corpora were used by only one reference each (see SWIFT-Review project file for references to each corpus).

From the 84 papers, at least 65 reported creating their own datasets. This included either curating a full dataset from scratch or labelling and using a smaller dataset in addition to a previously published benchmarking dataset. The high usage of own and custom datasets can in part

be explained by the heterogeneous characteristics of non-peer-reviewed data.

Peer-reviewed literature itself is commonly published in English. In contrast to that, information extracted from publicly available and grey literature sources is much more diverse. Social media posts, forum discussions or data from government agencies are often available in native languages of the authors who conduct the NLP research. We found examples for Indonesian,⁴⁵ Chinese,^{46,47} Croatian,⁴⁸ and other multilingual datasets including French or Latin.⁴⁹ This diversity in both data sources and in languages leads to a greater need to curate datasets on a project-by-project basis.

Currently, the biggest publicly available datasets include tweets and forum posts mostly in English, while availability of multilingual data is more limited. Datasets

are often small and include imbalances within the labelled classes, which in turn makes it difficult to train reliable and well-performing algorithms for automation.⁵⁰ Saha et al⁵⁰ also suggest that future research into using well-trained and evaluated methods for automatic translation into English is warranted. The performance of machine-translation and so-called multilingual zero-shot classification has improved greatly in recent years, but its application to data extraction for medical contexts has not been evaluated in great detail due to a lack of multilingual corpora.

Types of data

Adverse events ($n = 35$), followed by disease ($n = 29$), and drugs ($n = 18$) were the most common types of data addressed by the automation models in the included papers (see Figure 4). We tagged a total of 32 different types of data that are commonly used within health-related literature analyses. Some of these, such as the 14 papers categorised under ‘Technology and Trends’ can be of use to researchers implementing horizon scanning automation, while others such as ‘Symptom’ ($n = 16$) or ‘Intervention’ ($n = 14$) may be useful for a variety of literature review questions and methodologies.

Extent of automation

For each included paper, we tagged the extent of automation based on three options. These are explained in more detail below, and key aspects are further described in the glossary, which is given in Appendix B in Data S1. Table 6 summarises the results.

- Prioritisation and summarisation of evidence (lowest level of automation).
- Mining entities and sentences.
- Extraction and normalisation to standardised vocabularies (highest level of automation).

Prioritisation and summarisation of evidence. In the research context of this review, prioritisation and summarisation is very similar to the well-known NLP task of document-level classification and topic modelling. We found 63 papers that described the functionality of helping researchers to summarise, re-order or identify whole documents related to a health-related research question or task. This can be achieved by classifying whole tweets by content type or by identifying emerging technology trends within YouTube video captions or patents. For example, a tweet could be classified and prioritised as containing content about adverse events in general. This process helps to streamline the identification of relevant content by

presenting researchers with a pre-filtered set of likely-relevant research. Documents, as such, are not being data-extracted but rather pre-sorted and prepared for analysis. This process was included in the scope of this review paper because it is generally regarded as one of the most straightforward applications of AI and machine-learning, with a chance for high, reliable model performances. In the domain of screening peer-reviewed papers for systematic reviews, using AI in the prioritisation process is now widely recognised approach to save significant amounts of time in order to find relevant papers.^{23,26}

Out of the 63 papers, 25 added more value by combining the prioritisation step with the more specific task of mining entities or sentences, and 13 of those papers also covered the whole process of creating structured data by adding normalisation functionalities. In the following paragraphs, we describe the tasks of mining and normalisation in more detail, to give an overview of those processes and their potential added value. As part of this evidence map, we share a SWIFT-Review project that contains all included papers and the tags discussed in this section, such that readers can browse the papers in each category easily.

Mining entities and sentences. Here, we tagged papers if they described processes that lead to the targeted identification of shorter pieces of information in text, for example sentences, named entities, or relations between them. An entity could be a single word or short phrases of text belonging to a clearly defined class of things, such as the word ‘Aspirin’ being an entity of the class ‘Drug’.

Tasks related to data mining are harder than the prioritisation or summarisation, because they often require classification on a word-by-word basis and thus introduces a higher chance of errors or partly-correct identification of entities. The input text was usually natural language, in full texts or segmented into units such as sentences, abstracts, or paragraphs. In the included papers, 46 reported some form of mining functionalities within their text, mostly limited to named-entity recognition and not focussing on relation extraction. In practice, this leaves the user with selected pieces of text in a semi-structured form, because the resulting text is shorter and has class-assignments, such as ‘drug’ or ‘disease’. However, the mined text itself is just a subset of the original text, and therefore still present in the form of natural language. This natural language can carry variations in expressions that complicate automated synthesis of the data, thus still requiring human assistance and downstream manual work.

Extraction and normalisation. To create fully structured data from unstructured text, all mined text can be normalised to a structured vocabulary. For example, when normalising to MeSH terminology, mined text pieces such as ‘2-(Acetyloxy)benzoic Acid’, ‘Polopiryna’, or ‘acetylsalicylic acid’ would all be resolved to MeSH term ‘D001241: Aspirin’. In total, 17 papers described normalisation as part of their pipeline. The task of normalisation is harder than mining entities or sentences, because the core-classification task is not binary (i.e., not a choice between the decision drug/not-drug for a word) but rather a complex multi-class and sometimes multi-label case where the potential decision-space is as large as the vocabulary to normalise to. For example, when normalising to MeSH terms, there are more than 680,000 entry terms that can be chosen to normalise an entity to. This does not only create computational problems because of the large space of potential labels, but also problems in terms of ambiguity, non-covered vocabulary, and variations in specificity of the chosen concepts. In other words, one mined piece of text may correctly refer to one, more than one, or to no covered concept within a vocabulary. Whenever more than one correct concept applies, one might need to make the choice between less specific normalisations (i.e., high-level concepts in the MeSH tree) or more specific normalisations (i.e., the lowest-level finer-grained concepts). This increases not only the complexity of the classification task, but also makes it challenging to conduct a fair and comprehensive evaluation that is representative of future, unseen data that will be

seen by the system when it is deployed in practice. Furthermore, in practice, a correct normalisation requires correct named-entity recognition in the first place, thus escalating any errors made during downstream data processing. This accumulation of error during multiple classification-steps is a challenge that may be further reducing the amount of correctly normalised entities when tools and methods are used in practice.

Evaluation of algorithms

Evaluations were most commonly performed in a quantitative manner by using manually or distantly labelled gold-standard datasets. Most commonly, the process included the creation of a dataset by experts, and then splitting data randomly into training, validation, and test sets, to ensure that none of the data seen by an algorithm during the training-phase is used for evaluation. The process of splitting the data was either described in the papers, or authors described using published benchmark-datasets with predetermined splits. In line with scores frequently used to report automated data extraction results on peer-reviewed literature,¹³ the commonly used evaluation metrics of precision, recall, and F1 score were the most prevalent scores, with 19 papers reporting all three scores. F1 by itself was the most common score, reported in 47 papers, followed by precision ($n = 29$) and recall ($n = 26$). Accuracy was reported in $n = 9$ papers, MAP in $n = 4$, area-under-curve $n = 3$, specificity $n = 2$, one devised a new score and two papers used speed (see Figure 5).

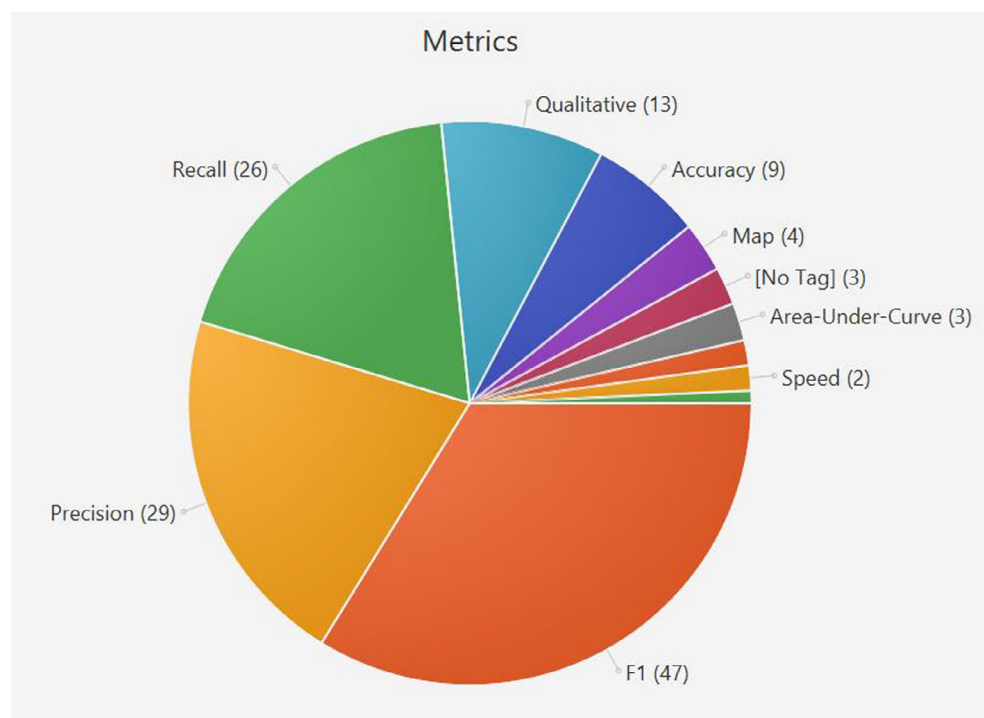


FIGURE 5 Level of automation in the included papers. One paper can include more than one metric for evaluation. [Colour figure can be viewed at wileyonlinelibrary.com]

Those scores were reported across different kinds of classification tasks, generally showing very good scores for straightforward tasks that include only document-level classification. For these binary classification tasks F1, precision, and recall scores higher than 0.9 is becoming more common.^{51–55} Scores decrease for harder classification tasks such as normalisation to controlled vocabularies, commonly ranging between 0.2 and 0.6 in precision, recall or F1. In part, this large variation between reported scores, spanning 0.2–0.6, can be seen for normalisation because the included papers used the same evaluation metrics but applied them in different ways. An example for this is the usage of relaxations such as counting a predicted answer as ‘correct’ if the true label was predicted within the top-N predictions, as opposed to only accepting one answer. Another relaxation method was to decrease the amount of potential labels to only include top-level categories.^{47,49,50,56}

Even when grouping and comparing classification scores for the three classification tasks separately, it is not straightforward to determine the best-performing algorithm within each category. Algorithm performance reported on a domain-specific dataset labelled by a group of researchers according to their own annotation guidelines may vary when the same algorithm is tested on completely new and therefore independent data labelled by different persons. A number of included papers used more than one dataset to evaluate their classifiers in parallel, showing differences between the evaluations scores of the same algorithm or architecture and therefore making it hard to estimate how each algorithm would perform in real-life, with potentially new or evolving data.^{57–59}

In total, 15 papers described qualitative or practical evaluations of their algorithms. In the most cases, the qualitative evaluation completely replaced quantitative analysis. Evaluations were conducted via case-studies and explorative analyses of large sets of unlabelled data. By applying the algorithm or proposed tools to a real-life dataset, authors discussed their perceived value of the automatically mined or extracted data. In the field of mining emerging health technologies, for example, it was a common approach to train Latent Dirichlet Distribution (LDA) topic model algorithms to identify dominant themes in large corpora of data.^{47,54,60–62} This approach is chosen because LDA is a generative unsupervised machine-learning approach, assigning a pre-defined number of topics to each document and using probability distributions across the vocabulary to assign words to topics.⁶³ The output of this algorithm is a set of unlabelled word-clouds with vocabulary that may have emergent semantic similarities when examined by a human. Those outputs are then qualitatively evaluated by picking emerging topics and discussing or visualising data.

Drawbacks for using topic-models such as LDA are that they require human interpretation, they are strongly influenced by training parameters (such as the pre-defined number of topics per document) and in the absence of a fully labelled dataset it is not possible to estimate sensitivity (i.e., to estimate the amount of missed important topics). However, barriers to conducting a full quantitative evaluation can include, for example, a lack of resources to create own labelled datasets where none are publicly available. For this reason, researchers are opting for such unsupervised algorithms to create value, without investing large amounts of resources.⁶⁴

Practical evaluations of tools: Evaluations done for tools, in the absence of labelled data, include assessments of efficiency within the workflow, by doing a direct comparison between time-taken by humans to complete a task versus AI-supported humans receiving automatically extracted or likely relevant data first. Zhao et al,³² for example, found that using their unnamed AI tool in explorative analysis of biomedical video content increased the speed of finding relevant evidence by 3.7 times, and that users were able to answer content-based questions more accurately. Keeling et al³⁴ qualitatively compared their search tool PHIS with Google, applying the ‘Critical Incident Technique’ by asking users to keep notes of situations where the tools were effective or ineffective, and conducting retrospective semi-structured interviews.

3.2.3 | Practical challenges and research gaps that constitute barriers to the development and deployment of data extraction tools and methods

Heterogeneity and transferability within the data

Heterogeneity within the data was discussed as a barrier to the overall process by Sofean and Aras.⁶¹ They focused on patent mining and noted that within patent documents there are different kinds of text such as the main text and metadata, including information about the inventors, institutions, and timelines. The challenge of automating data extraction from patents exceeds the boundaries of NLP, because they include potentially valuable information in forms other than text, for examples as drawings or schemes.⁶¹ Turner et al⁴² described heterogeneity within 320 analysed grey literature documents. This included the challenges caused by different document types such as HTML or PDF, different content types such as text or figures or tables, and a general broadness and inconsistency of topics and subject matters.⁴² Other features of unstructured data that cause challenges are colloquialisms, abbreviations, spelling

errors, and other variations that appear in natural language.^{41,56,65,66}

Chee et al⁶⁵ describe heterogeneity not within a data source, but between them. Due to different text lengths and/or languages it becomes hard to achieve knowledge- or domain-transfer between sources such as Twitter and online forum entries. This creates a need for developing separate datasets and classifiers for extracting the same type of information, for example drugs, from these heterogeneous sources.^{64,65,67}

Complexity, noise, and ambiguity within the data

Complexity is a challenge mostly described within papers that attempted the normalisation of extracted data, due to the large space of possible terms to map to.⁶⁸ Another factor adding on to complexity of tasks is unstructured or irrelevant background information within datasets, or a lack of context within short texts such as tweets.^{68,69} Ambiguity also increases complexity, for example when drug name can have multiple synonyms, trade names, or multiple correct labels.^{39,56} Noise is a concept that generally refers to data being unreliable due to their unstructured and naturally expressed form, thus causing errors both while labelling gold-standard data and when processing and predicting on new data.^{31,52}

Sparsity or imbalance within the data

ML or neural networks architectures require annotated data during the learning process. When training classifiers for forum posts to identify drugs, Chee et al⁶⁵ noted that some drugs did not have enough mentions in posts in order to train and evaluate robust classifiers, while Arnold et al⁵⁷ also describe rare entities and unseen data as a problem. Imbalance in the data, when there are more positive training examples for certain types of information, can be another issue that leads to performance drops in under-represented classes.^{70,71}

Scalability

Data extraction methods are trained and evaluated on benchmark datasets. However, when deploying them for practical use in real-world scenarios, the amount of data that needs to be processed increases, causing processing times of multiple hours or days and necessitating the use of high-quality hardware and analytics platforms.^{34,36,40} Ul Haq et al³³ discuss that this is a complex task because classification accuracy, time, and versatility need to scale in parallel with the real-world tasks that a system is applied to solve.

Corpus availability and cost to generate annotated data

Multiple publications described a lack of publicly corpora and benchmark datasets. This is a commonly described issue for different types of data, including patents⁷² or

social media.⁷³ It is also mentioned specifically in relation to text extraction and normalisation to standardised vocabularies such as the UMLS.⁶⁸

In the absence of publicly available corpora, researchers are forced to spend money and resources to create their own customised datasets,⁶⁵ which can lead to small datasets with limited usefulness, thus creating the need to adapt models to maximise gain in situations where better performances could be achieved.⁷⁴

When using labelled gold-standard social-media corpora from sources such as Twitter, copyright, and data-availability were described as a challenge.⁷⁵ Twitter datasets can shrink as tweets become unavailable over time, thus reducing reproducibility and comparability of results obtained at different points in time when using the same corpus.^{56,75,76}

However, there exists a vast amount of secondary research that has utilised automated data extraction in practice. These tend to be AI project-specific tools extracting data to create very targeted and topic-specific intelligence using bespoke methods that are generally not re-usable beyond the original research project. Due to the vast amount of topic-specific automation we did not include these papers in the full-text analysis, but rather tagged them by topic (e.g., mental health, COVID-19) and data source (Twitter, news). This supplementary evidence map includes 318 papers. A description of our findings, along with figures and a heat-map is provided in Appendix C in Data S1, and an interactive version giving access to the papers and their data tags is provided within the SWIFT-Review project in the [Supporting Information](#).⁶

4 | DISCUSSION

The surge in published literature and the dearth of intelligence available is driving the need for more innovative methods to deliver timely secondary research. Fully or semi-automated data extraction may offer a means to make both unstructured and structured data more accessible to those undertaking this type of research. However, conducting any secondary research projects is time-intensive, and often projects themselves are time-sensitive. Therefore, it might not always be feasible to include evidence from lower-quality, grey literature or soft data sources. Fully or semi-automated data extraction can be a way forward, to make unstructured data accessible and facilitate integration into review workflows. However, this works only if data can be automatically identified and extracted using a targeted, well evaluated and evidence-based approach. Further, it remains important that the data being used, whether it is from

RCTs or Twitter, are pertinent to the question being asked and the decision being made.

We included 7 end-user tools and 76 published methods papers for data extraction of grey literature and non-peer-reviewed data in this mapping review. There is a broad range of secondary health-related research that could benefit from using grey literature and softer data. Horizon scanning is one use-case, because it utilises timely, soft sources of information to detect signals of future trends in research and technology.⁶ Similar use-cases for softer and automatically extracted data include the identification of future research topics and protocol formulation. Another potential use-case is the inclusion of rapid analysis of these non-peer-reviewed data sources in the discussion section of systematic reviews, where impact on patients and practitioners, impact on health-care systems, research gaps in clinical trials or recommendations for the future are discussed.

4.1 | Discussion in the context of related literature reviews

Correia et al¹⁶ published a narrative review of recent work on data mining in social media content analysis. They discuss papers on automation in the domains of pharmacovigilance and sentiment analysis, most commonly targeting specific drugs and their adverse events, or mental health research questions. These findings correspond to our mapping of the topic-specific literature (see Appendix C in Data S1), we mapped mental health and sentiment analysis within the top-3 applications. We picked up COVID-19 within the top-3; this is not represented within Correia et al¹⁶ due to their publication date in May 2020. They discuss limitations specific to social-media data, such as limitations of reliability of the data when users build online-personas, limitations related to bot-content, limitations when sampling data for analysis, and caveats that people posting online are a small selection from a wider publication and thus samples may not be representative.¹⁶ A living systematic review of automated data extraction from the highly related field of peer-reviewed health literature currently includes 76 papers, indicating fast-paced advancements in the areas of automatic extraction, normalisation, relation extraction and text summarisation.¹³ Their main conclusions are similar to the findings of this review, citing low availability of end-user tools among many published methods of data extraction leading to slow uptake of automation methods in practice, low comparability between evaluation results, and high duplication of research efforts.¹³

4.2 | Discussion of important features of extraction tools

Most included papers described methods for extraction of data, with potential features that might be beneficial for future tools. The identified tools discussed accessibility (i.e., as web-application), bulk processing of text, automatically updating data from the web, automatic query expansion, and visualisations as main features. Barriers to integration with other downstream tools was identified as a research gap.

4.3 | Discussion of the level of support given by tools and methods

To encourage practical use of the included tools and methods, their underlying NLP methods need to be accessible in the form of usable tools, connected to online data-sources for automatic information retrieval, and well validated. In summary, three different types of evaluation were described in the included publications to validate models, each applied as required by the task and research context:

1. A direct model performance validation, where the proposed model is compared with other published models or algorithms that were trained using the same dataset, ideally with the same train/validation set splits.
2. An adaptability validation, where the proposed model's evaluation scores are compared with the same model's scores across different independent datasets that often, but not necessarily, fit the same domain but have different characteristics such as data source, annotation guidelines, or topic-distributions.
3. A practical validation, where the model is used to make predictions on real-life, unlabelled data. This evaluation can be qualitative, in terms of perceived usefulness or trustworthiness of the system as part of a case-study, or comparative in terms of time-saved during screening and data extraction.

4.4 | Discussion of practical challenges, research gaps, and caveats as described within the included papers

In the following section, we discuss the broader implications of our analysis, focussing on the limitations associated with using automated data extraction tools and methods in real-world scenarios. These limitations include both the point-of-view of the tool providers, in terms of challenges related to the deployment of usable tools, as well as general challenges caveats relating to

user's lack of trust and further research that is needed when integrating reliable and usable automation in data extraction into real-world research projects.

Natsiavas et al³⁹ described a method and tool design to be used by clinicians at the point of care; noting that integration into already established workflows can be challenging, due to already established routines and information overflow for the clinicians. They noted that having normalised data, facilitating data-sharing, and implementing continuous updates would be helpful, but acknowledge that those features are hard to implement.³⁹

In the real world, tools need to be accessible to users. Costs need to be calculated for hardware and providing computational resources and servers for deployment. This is challenging because it can make it expensive for tools to be live and accessible.⁶⁷

Another important issue that prevents the usage of automation methods is a lack of trust in the reliability of tools and methods, concerning for example trust into the sources of the information or a lack of high-performing classifiers that can provide adequate performance.^{30,34,71} Many of the included papers used their own datasets for training and evaluation, or completed their evaluation using different evaluation metrics. This severely limits the comparability of approaches and cannot provide us with definite answers on how trustworthy or reliable tools are. Whenever methods were tested in real-world scenarios, these tests were usually small and unmeaningful as the process of using the automation approach was not directly compared with a fully manual analyses on the same dataset. Outcomes such as time-saved, or number of relevant records discovered by each method, were rarely assessed.

Non-scientific or grey literature data from social media, patents, or similar sources are often expressed in languages other than English. This means that problems around sparsity, data imbalance, and lack of availability are exacerbated, as they make it harder to obtain good representations of the language and tasks that need to be achieved. A potential solution to this problem is the usage of automatic translation software. With the advent of neural networks in NLP the performance of machine-translation algorithms has steadily improved over the past years,⁷⁷ and a selection of free or paid-for APIs such as DeepL⁷ and Google Translate⁸⁹ are available to process information in various formats.

A potential practical challenge we noted is the evaluation of a tool or model on data that has previously been seen in training. This issue should not arise when training and evaluating on one dataset that has been correctly split into train and test data, but it may arise when multiple datasets are created from the same source and then

subsequently used as additional evaluation sets. For example, [ClinicalTrials.gov](https://clinicaltrials.gov) is a frequent source of data described in multiple papers and datasets^{21,78–81} and thus caution needs to be exercised when using or evaluating across datasets that are available from related research projects.

4.5 | Recommendations for tool development

In summary, researchers or companies looking to develop automated data extraction tools should consider the financial implications for tool development and deployment, and the scalability of their automation methods to estimate hardware needs and running costs in the long-term. They should also carefully assess user needs and interoperability of the proposed tool with other down or upstream tools used in literature analysis or in clinical practice. These considerations determine the complexity of their proposed tool and should be key to the planning, execution, and evaluation stages of the final tool.

Unfortunately, this does not guarantee user-acceptance, and significant risks remain when investing research time and money into tool development. A lack of trust in the tool and/or its underlying automation methods may still lead to a lower-than-expected uptake, at which point the effort of maintaining tools is too high and may not be worth it.

To increase acceptance, transparent large-scale testing and comparisons on different datasets may be needed, in conjunction with early engagement with the research community during the design phase and later via publications in peer-reviewed journals. As discussed in the previous sections, a comprehensive evaluation includes comparing and contrasting one automation model with other models using the same datasets, applying the model to multiple datasets with different characteristics to simulate different real-world projects, and running large-scale practical evaluations on real-world projects without any pre-defined gold-standard data, to measure outcomes important to researchers who conduct literature reviews. These outcomes could be time-savings between automation-supported and manual processes, number of records missed or gained through automation, and usability and integrability of the tool into established workflows and methodologies. None of the included papers mentioned marketing, to increase awareness and public knowledge about tools.

4.6 | Limitations

The scope of this mapping review is intentionally broad. A vast amount of literature exists in the intersection between automatic data extraction and health-related evidence/data in the public domain. We aimed to be systematic in capturing the relevant literature during the search but limited the inclusion criteria to papers that extract general-purpose text (as opposed to including topic-specific analyses). To mitigate this limitation, we have separated and tagged all topic-specific papers as part of a separate evidence map based on title/abstract information of 318 papers. We presented an abbreviated version of these results within Appendix C in Data S1.

We extracted evaluation scores for included papers and discussed model performances for tasks of differing complexity but did not directly compare the performance of any methods discussed in this review. We avoided drawing conclusions on the 'best' tools or methods available; this was not our aim. However, had we sought to do this feasibility of such an exercise would have been hindered by the usage of different datasets and different methods of evaluations across the papers. Further in-depth case-studies that include implementation and direct comparisons of some of these methods are reserved for future work.

5 | CONCLUSION

This review summarises current knowledge about functionalities, data sources, and performance of published methods and tools to automate data extraction of grey literature and soft data related to healthcare. We performed a detailed analysis of key strengths and weaknesses of 7 end-user tools and 76 methods papers, and the level of support they provide. We collected information about barriers in implementing automation in practice, and a summary of caveats and experiences from using automatically mined and extracted data in real world projects. Overall availability of code, data, and implementation of methods into accessible end-user tools was poor, suggesting that the field of automating grey-literature mining suffers from high duplication of research efforts, and at the same time low uptake of the few tools and methods that are available.

5.1 | Highlights

This is the first review of automated data extraction from health-related grey literature and soft data; to

automate horizon scans, HTAs, evidence maps or other secondary literature reviews. It includes 84 tools and methods papers mining information from health-related news, patents, websites, trial registers, fora, or social media.

We discuss relevant end-user features of tools, types of extracted data and text such as 'disease' or 'outcome', evaluation metrics and results, practical implications of usage, research gaps and barriers to development and deployment of automation methods in this field in practice.

This review provides a detailed insight into automated classification, mining, and normalisation of data from grey literature and soft data. We tagged, mapped, and shared all results to enable both data scientists and researcher with health-related research background to easily filter and access all included papers.

AUTHOR CONTRIBUTIONS

Lena Schmidt: Conceptualization; methodology; software; data curation; investigation; visualization; writing – original draft. **Saleh Mohamed:** Methodology; validation; writing – review and editing; data curation. **Nick Meader:** Methodology; validation; writing – review and editing; supervision; data curation. **Jaume Bacardit:** Methodology; writing – review and editing; supervision. **Dawn Craig:** Conceptualization; methodology; writing – review and editing; supervision; funding acquisition.

ACKNOWLEDGMENTS

We would like to thank Chris Marshall for giving very valuable feedback on the methodology of this paper during protocol development and for providing edits and feedback to an earlier draft. We would also like to thank Fiona Beyer for her feedback on the search strategies and her recommendations regarding the databases to search for this review.

FUNDING INFORMATION

This project is funded by the National Institute for Health and Care Research (NIHR) [HSRIC-2016-10009/Innovation Observatory]. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All references included on full text and for the separate map of titles and abstracts, together with their tags and extracted data, are available in the form of SWIFT-

Review project files in the [Supporting Information](#). These .stp files can be opened using: <https://www.sciome.com/swift-review>. We also created a repository on Harvard Dataverse to host the project files, available here: Schmidt, Lena, 2023, 'Automated data extraction of unstructured and grey literature data in health research: a mapping review of the current research literature', <https://doi.org/10.7910/DVN/7N2YWZ>.

ORCID

Lena Schmidt  <https://orcid.org/0000-0003-0709-8226>

ENDNOTES

- <https://www.python.org/doc/versions/>
- <https://www.java.com/releases/>
- <https://cran.r-project.org/web/packages/medrxivr/index.html>
- <https://www.sciome.com/swift-review/>
- <https://live.european-language-grid.eu/catalogue/corpus/5090>
- [10.7910/DVN/7N2YWZ](https://doi.org/10.7910/DVN/7N2YWZ)
- <https://www.deepl.com/pro-api?cta=header-pro-api>
- <https://cloud.google.com/translate>
- <https://py-googletrans.readthedocs.io/en/latest/>

REFERENCES

- Blumenfeld P, Pfeffer RM, Symon Z, et al. The lag time in initiating clinical testing of new drugs in combination with radiation therapy, a significant barrier to progress? *Br J Cancer*. 2014;111(7):1305-1309. doi:10.1038/bjc.2014.448
- 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):510-520. doi:10.1258/jrsm.2011.110180
- Van Norman GA. Drugs, devices, and the FDA: part 2: an overview of approval processes: FDA approval of medical devices. *JACC: Basic Transl Sci*. 2016;1(4):277-287. doi:10.1016/j.jacbts.2016.03.009
- DeYoung J, Beltagy I, van Zuylen M, Kuehl B, Wang LL. MS²: a dataset for multi-document summarization of medical studies. *ArXiv*. 2021. doi:10.48550/arXiv.2104.06486
- Goodman CS, Church F. HTA 101 Introduction to health technology assessment. 2004.
- Hines P, Hiu Yu L, Guy RH, Brand A, Papaluca-Amati M. Scanning the horizon: a systematic literature review of methodologies. *BMJ Open*. 2019;9(5):e026764. doi:10.1136/bmjopen-2018-026764
- Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev*. 2007;2007(2):Mr000010. doi:10.1002/14651858.MR000010.pub3
- Paez A. Grey literature: an important resource in systematic reviews. *J Evid Based Med*. 2017;10:233-240. doi:10.1111/jebm.12265
- Lefebvre C, Glanville J, Briscoe S, et al. Searching for and selecting studies. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons; 2019:67-107. doi:10.1002/9781119536604.ch4
- Singh L, Bode L, Budak C, Kawintiranon K, Padden C, Vraga E. Understanding high- and low-quality URL sharing on COVID-19 twitter streams. *J Comput Soc Sci*. 2020;3(2):343-366. doi:10.1007/s42001-020-00093-6
- WHO. *Health Technology Assessment Survey 2020/21—Main Findings*. WHO; 2021 <https://www.who.int/data/stories/health-technology-assessment-a-visual-summary>
- Lauvrak V, Arentz-Hansen H, Di Bidino R. Recommendations for horizon scanning, topic identification, selection and prioritisation for European cooperation on health technology assessment. 2020 <https://www.eunetha.eu/wp-content/uploads/2020/04/200305-EU-netHTA-WP4-Deliverable-4.10-TISP-recommendations-final-version-1.pdf>
- Schmidt L, Olorisade BK, McGuinness LA, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: a living systematic review. *F1000Research*. 2021;10:401.
- Acosta-Urigüen M-I, Arias B, Orellana M. Text Mining Techniques Implemented to Extract Data from Transit Events in Twitter: A Systematic Literature Review. In: Rodriguez G, Morales ER, Fonseca C, et al., eds. *Information and Communication Technologies*. Springer; 2020.
- Viviani M, Pasi G. Credibility in social media: opinions, news, and health information—a survey. *WIREs Data Min Knowl Discov*. 2017;7(5):e1209. doi:10.1002/widm.1209
- Correia RB, Wood IB, Bollen J, Rocha LM. Mining social media data for biomedical signals and health-related behavior. *Annu Rev Biomed Data Sci*. 2020;3:433-458. doi:10.1146/annurev-biodatasci-030320-040844
- Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018;25(10):1419-1428. doi:10.1093/jamia/ocy068
- Miller DM, Shalhout SZ. GENETEX—a GENomics report TEXT mining R package and shiny application designed to capture real-world clinico-genomic data. *JAMIA Open*. 2021;4(3):ooab082. doi:10.1093/jamiaopen/ooab082
- Cabanac G, Oikonomidi T, Boutron I. Day-to-day discovery of preprint-publication links. *Scientometrics*. 2021;126(6):5285-5304. doi:10.1007/s11192-021-03900-7
- Liu S, Bourgeois FT, Dunn AG. Identifying unreported links between ClinicalTrials.gov trial registrations and their published results. *Res Synth Meth*. 2022;13(3):342-352. doi:10.1002/jrsm.1545
- Smalheiser NR, Holt AW. A web-based tool for automatically linking clinical trials to their publications. *J Am Med Inform Assoc*. 2022;29:822-830. doi:10.1093/jamia/ocab290
- Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev*. 2015;4(1):78. doi:10.1186/s13643-015-0066-7
- O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev*. 2015;4(5). doi:10.1186/2046-4053-4-5
- McGuinness LA, Schmidt L. Medrxivr: accessing and searching medRxiv and bioRxiv preprint data in R. *J Open Source Softw*. 2020;5:2651. doi:10.21105/joss.02651

25. Howard BE, Phillips J, Miller K, et al. SWIFT-review: a text-mining workbench for systematic review. *Syst Rev.* 2016;5(1):87. doi:10.1186/s13643-016-0263-z
26. Howard B, Phillips J, Tandon A, et al. SWIFT-active screener: accelerated document screening through active learning and integrated recall estimation. *Environ Int.* 2020;138:105623. doi:10.3389/fdgth.2020.592237
27. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst Rev.* 2022;18(2):e1230. doi:10.1002/cl2.1230
28. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. doi:10.1136/bmj.n71
29. Rabatel J, Arsevska E, Roche M. PADI-web corpus: labeled textual data in animal health domain. *Data Brief.* 2019;22:643-646. doi:10.1016/j.dib.2018.12.063
30. Valentin S, Arsevska E, Rabatel J, et al. PADI-web 3.0: a new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health.* 2021;13:100357. doi:10.1016/j.onehlt.2021.100357
31. Pasche E, Gobeill J, Teodoro D, et al. A user-friendly tool for medical-related patent retrieval. *Stud Health Technol Inform.* 2012;174:121-125.
32. Zhao B, Xu S, Lin S, Luo X, Duan L. A new visual navigation system for exploring biomedical Open Educational Resource (OER) videos. *J Am Med Inform Assoc.* 2016;23(e1):e34-e41. doi:10.1093/jamia/ocv123
33. Ul Haq H, Kocaman V, Talby D. Mining adverse drug reactions from unstructured mediums at scale. *ArXiv.* 2022. doi:10.48550/arXiv.2201.01405
34. Keeling JW, Turner AM, Allen EE, et al. Development and evaluation of a prototype search engine to meet public health information needs. *AMIA Annu Symp Proc.* 2011;2011:693-700.
35. Avasarala V, Bonissone P. iPresage: An innovative patent landscaping tool. *2012 IEEE Congress on Evolutionary Computation, Brisbane, QLD, Australia, 2012*, pp. 1-7. doi:10.1109/CEC.2012.6256503.
36. Kravets AG, Korobkin DM, Dykov MA. E-Patent Examiner: Two-Steps Approach for Patents Prior-Art Retrieval. *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Corfu, Greece, 2015, pp. 1-6.
37. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics Avignon.* 2012.
38. Hariprasad S, Xue-wen C, Bo L. Ontology-based visualization of healthcare data mined from online healthcare forums. *2015 International Conference on Healthcare Informatics*, Dallas, TX, 2015, pp. 325-334. doi:10.1109/ICHI.2015.46
39. Natsiavas P, Jaulent MC, Koutkias V. A knowledge-based platform for assessing potential adverse drug reactions at the point of care: user requirements and design. *Stud Health Technol Inform.* 2019;264:1007-1011. doi:10.3233/shti190376
40. Tafti AP, Badger J, LaRose E, et al. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR Med Inform.* 2017;5(4):e51. doi:10.2196/medinform.9170
41. Lee K, Uzuner Ö. Normalizing adverse events using recurrent neural networks with attention. *AMIA Jt Summits Transl Sci Proc.* 2020;2020:345-354.
42. Turner AM, Liddy ED, Bradley J, Wheatley JA. Modeling public health interventions for improved access to the gray literature. *J Med Libr Assoc.* 2005;93(4):487-494.
43. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: a corpus of adverse drug event annotations. *J Biomed Inform.* 2015;55:73-81. doi:10.1016/j.jbi.2015.03.010
44. Alvaro N, Miyao Y, Collier N. TwiMed: twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill.* 2017;3(2):e24. doi:10.2196/publichealth.6396
45. Halim C, Wicaksono AF, Adriani M. Extracting Disease-Symptom Relationships from Health Question and Answer Forum. *2017 International Conference on Asian Language Processing (IALP)*, Singapore, 2017, pp. 87-90.
46. Chen Y, Zhou C, Li T, et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. *J Biomed Inform.* 2019;96:103252. doi:10.1016/j.jbi.2019.103252
47. Zhao S, Jiang M, Yuan Q, Qin B, Liu T, Zhai C. ContextCare: Incorporating Contextual Information Networks to Representation Learning on Medical Forum Data. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press. 2017, pp. 3497-3503.
48. Kocijan K, Kurolt S, Mijić L. Building croatian medical dictionary from medical corpus. *Rasprave Inst Za Hrvatski Jezik i Jezikoslovlje.* 2020;46(2):765-782. doi:10.31724/RIHJJ.46.2.17
49. Grabar N, Hamon T. Automatic Extraction of Layman Names for Technical Medical Terms. *2014 IEEE International Conference on Healthcare Informatics*, Verona, Italy, 2014, pp. 310-319.
50. Saha S, Das S, Khurana P, Srihari R. Autobots Ensemble: Identifying and Extracting Adverse Drug Reaction from Tweets Using Transformer Based Pipelines. 2020 <https://aclanthology.org/2020.smm4h-1.16>
51. Atal I, Zeitoun JD, Névéol A, Ravaud P, Porcher R, Trinquart L. Automatic classification of registered clinical trials towards the global burden of diseases taxonomy of diseases and injuries. *BMC Bioinformatics.* 2016;17(1):392. doi:10.1186/s12859-016-1247-7
52. Ellendorff T, Cornelius J, Gordon H, Colic N, Rinaldi F. UZH@SMM4H: System Descriptions. 2018. doi:10.18653/v1/W18-5916
53. Fan B, Fan W, Smith C, Garner H. Adverse drug event detection and extraction from open data: a deep learning approach. *Inf Process Manage.* 2020;57(1):102131. doi:10.1016/j.ipm.2019.102131
54. Guo H, Na X, Li J. Automatically identifying topics of consumer health questions in Chinese. *Stud Health Technol Inform.* 2017;245:388-392.
55. Rezaei Z, Ebrahimpour-Komleh H, Eslami B, Chavoshinejad R, Totonchi M. Adverse drug reaction detection in social media by Deepm learning methods. *Cell J.* 2020;22(3):319-324. doi:10.22074/cellj.2020.6615
56. Magge A, O'Connor K, Scotch M, Gonzalez-Hernandez G. SEED: symptom extraction from English social media posts using deep learning and transfer learning. *medRxiv.* 2021a. doi:10.1101/2021.02.09.21251454

57. Arnold S, Van Aken B, Grundmann P, Gers FA, Löser A. Learning Contextualized Document Representations for Healthcare Answer Retrieval. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, 2020, pp. 1332–1343.
58. Guo Y, Ge Y, Yang YC, Al-Garadi MA, Sarker A. Comparison of pretraining models and strategies for health-related social media text classification. *Healthcare*. 2022. doi:10.1101/2021.09.28.21264253
59. Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *J Biomed Inform*. 2016;62:148-158. doi:10.1016/j.jbi.2016.06.007
60. Daniel C, Dutta K. *Automated Generation of Latent Topics on Emerging Technologies from YouTube Video Content*. AIS Electronic Library (AISeL). 2018.
61. Sofean M, Aras H. Technological Areas Detection and Clustering for Large-Scale of Patent Texts. In *The International Conference on Big Data Analytics, Data Mining and Computational Intelligence 2018 (BigDaCI)*. 2018, Madrid, Spain; iadis digital library. 2018.
62. Zhou Y, Dong F, Liu Y, Ran L. A deep learning framework to early identify emerging technologies in large-scale outlier patents: an empirical study of CNC machine tool. *Scientometrics*. 2021;126(2):969-994. doi:10.1007/s11192-020-03797-8
63. Blei DM, Ng AY, Jordan MJ. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
64. Rastegar-Mojarad M, Liu H, Nambisan P. Using social media data to identify potential candidates for drug repurposing: a feasibility study. *JMIR Res Protoc*. 2016;5(2):e121. doi:10.2196/resprot.5621
65. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. *AMIA Annu Symp Proc*. 2011; 2011:217-226. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84863556111&partnerID=40&md5=72d17f7324c0865344858656d85b7506>
66. Woo HG, Yeom J, Lee C. Screening early stage ideas in technology development processes: a text mining and k-nearest neighbours approach using patent information. *Technol Anal Strat Manage*. 2019;31(5):532-545. doi:10.1080/09537325.2018.1523386
67. Jimeno-Yepes A, MacKinlay A, Han B, Chen Q. Identifying diseases, drugs, and symptoms in Twitter. *Stud Health Technol Inform*. 2015;216:643-647.
68. Batbaatar E, Ryu KH. Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. *Int J Environ Res Public Health*. 2019;16(19):3628. doi:10.3390/ijerph16193628
69. Gao J, Liu N, Lawley M, Hu X. An interpretable classification framework for information extraction from online healthcare forums. *J Healthc Eng*. 2017;2017:2460174. doi:10.1155/2017/2460174
70. Dai HJ, Wang CK. Classifying adverse drug reactions from imbalanced Twitter data. *Int J Med Inform*. 2019;129:122-132. doi:10.1016/j.ijmedinf.2019.05.017
71. Magge A, Tutubalina E, Miftahutdinov Z, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *J Am Med Inform Assoc*. 2021b;28(10):2184-2192. doi:10.1093/jamia/ocab114
72. Krishnan A, Cardenas AF, Springer D. Search for Patents Using Treatment and Causal Relationships. In *Proceedings of the 3rd international workshop on Patent information retrieval (PaIR '10)*. Association for Computing Machinery, New York, NY. 2010.
73. Yang M, Wang X, Kiang M. Identification of Consumer Adverse Drug Reaction Messages on Social Media. *PACIS 2013 Proceedings*. 2013, pp. 193.
74. Shen C, Lin H, Li Z, Chu Y, Yang Z. A Graph-Boosted Framework for Adverse Drug Event Detection on Twitter. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Seoul, Korea (South) 2020, pp. 1129–1131.
75. Zhang T, Lin H, Xu B, Yang L, Wang J, Duan X. Adversarial neural network with sentiment-aware attention for detecting adverse drug reactions. *J Biomed Inform*. 2021;123:103896. doi:10.1016/j.jbi.2021.103896
76. Karisani P, Ho J, Agichtein E. Domain-Guided Task Decomposition with Self-Training for Detecting Personal Events in Social Media. 2020 <https://export.arxiv.org/abs/2004.10201>
77. Wang H, Wu H, He Z, Huang L, Church KW. Progress in machine translation. *Engineering*. 2022;18:143-153. doi:10.1016/j.eng.2021.03.023
78. Goodwin TR, Skinner MA, Harabagiu SM. Automatically linking registered clinical trials to their published results with deep highway networks. *AMIA Jt Summits Transl Sci Proc*. 2018; 2017:54-63.
79. Patel CO, Cimino JJ. Semantic query generation from eligibility criteria in clinical trials. *AMIA Annu Symp Proc*. 2007;1070.
80. Pradhan R, Hoaglin DC, Cornell M, Liu W, Wang V, Yu H. Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses. *J Clin Epidemiol*. 2019;105: 92-100. doi:10.1016/j.jclinepi.2018.08.023
81. Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, Bian J, He Z. Transformer-Based Named Entity Recognition for Parsing Clinical Trial Eligibility Criteria. *ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2021, p. 49. doi:10.1145/3459930.3469560

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schmidt L, Mohamed S, Meader N, Bacardit J, Craig D. Automated data analysis of unstructured grey literature in health research: A mapping review. *Res Syn Meth*. 2023; 1-20. doi:10.1002/jrsm.1692