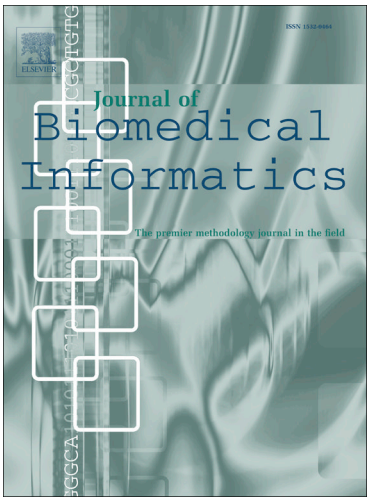


Patient data discovery platforms as enablers of biomedical and translational research: a systematic review

Alina Trifan, José Luís Oliveira

PII: S1532-0464(19)30072-3
DOI: https://doi.org/10.1016/j.jbi.2019.103154
Article Number: 103154
Reference: YJBIN 103154

To appear in: *Journal of Biomedical Informatics*

Received Date: 7 December 2018
Revised Date: 15 March 2019
Accepted Date: 18 March 2019

Please cite this article as: Trifan, A., Oliveira, J.L., Patient data discovery platforms as enablers of biomedical and translational research: a systematic review, *Journal of Biomedical Informatics* (2019), doi: https://doi.org/10.1016/j.jbi.2019.103154

Patient data discovery platforms as enablers of biomedical and translational research: a systematic review

Alina Trifan and José Luís Oliveira

IEETA/DETI, University of Aveiro, Portugal

Abstract

Background The global shift from paper health records to electronic ones has led to an impressive growth of biomedical digital data along the past two decades. Exploring and extracting knowledge from these data has the potential to enhance translational research and lead to positive outcomes for the population's health and healthcare.

Objective The aim of this study was to conduct a systematic review to identify software platforms that enable discovery, secondary use and interoperability of biomedical data. Additionally, we aim evaluating the identified solutions in terms of clinical interest and main healthcare-related outcomes.

Methods A systematic search of the scientific literature published and indexed in Pubmed between January 2014 and September 2018 was performed. Inclusion criteria were as follows: relevance for the topic of biomedical data discovery, English language, and free full text. To increase the recall, we developed a semi-automatic and incremental methodology to retrieve articles that cite one or more of the previous set.

Results A total number of 500 candidate papers were retrieved through this methodology. Of these, 85 were eligible for abstract assessment. Finally, 37 studies qualified for a full-text review, and 20 provided enough information for the study objectives.

Conclusions This study revealed that biomedical discovery platforms are both a current necessity and a significantly innovative agent in the area of healthcare. The outcomes that were identified, in terms of scientific publica-

Email address: {alina.trifan, jlo}@ua.pt (Alina Trifan and José Luís Oliveira)

tions, clinical studies and research collaborations stand as evidence.

Keywords: biomedical data discovery, discovery platforms, data interoperability, secondary use of data, translational research

1. Introduction

The wide adoption of digital records among health institutions across the world has been an important factor in shaping both current medical practices and clinical research. Electronic medical records (EMR), or Electronic Health Records (EHR), have successively replaced paper-based medical records, allowing the storage, retrieval and modification of clinical registers through digital means. They include routinely gathered clinical data, patient demographic data, laboratory results, radiology and pharmaceutical records, statistic and administrative data, as well as patient-centred, such as data coming from self-monitoring devices. With the increasing use of medical digital records, we are witnessing a shift in the complexity of the overall goal of medical institutions and physicians, which is to ultimately provide better healthcare. Nowadays the greatest difficulty is no longer to acquire data, but rather to manage it and extrapolate knowledge from it. The volume, speed and heterogeneity at which medical data is produced expose the so called Big Data era within healthcare [1]. Health Big Data (HBD) covers more than just a very large amount of data or a large number of data sources. It also takes into consideration the complexity, challenges and new opportunities presented by the combined analysis of data and their secondary use [2]. Now more than ever, there is a significant potential for the reuse of HBD for research [3]. Secondary use of health data has the potential to expand knowledge about diseases and their appropriate treatments, generate new understanding about the efficiency of healthcare systems, and fuel new discoveries that can eventually lead to a more personalized healthcare [4, 5, 6].

The integration and reuse of these huge amounts of data can impact clinical decisions, pharmaceutical discoveries, disease monitoring and the way the populations healthcare is provided globally. Storing data for future reuse and reference has been a critical factor in the success of modern biomedical sciences [7]. While the storing of electronic health data is done nowadays intrinsically and ubiquitously in more and more healthcare facilities, with many countries already relying fully on digital records, its reuse is still a delicate process and not at all straightforward. In order for data to be reused,

first it has to be discovered. Finding a dataset for a study can be burdensome due to the need to search individual repositories, read numerous publications and ultimately contact data owners or publication authors on an individual basis. Electronic health data discovery is raising the interest within the research community, due to the possibility to share and to study large datasets. Data discovery solutions usually focus, on one hand, on providing researchers with an overview of existent and accessible datasets and on the other hand, on giving data owners the possibility to communicate the existence of data, without necessarily fully exposing it, all in the same place. Translational research indicates a need for data discovery platforms that can stage and disseminate data in a readily accessible form [3]. Sharing health digital datasets for secondary use allows carrying out research studies with minor costs by leveraging existing data and achieving a better use of resources invested in research [8].

One important factor that impacts the quality of the data reuse process is its interoperability. Ideally, biomedical data would be universally stored following the same predefined set of rules and protocols. While efforts are being conducted in this direction, through the definition and adoption of ontologies and standard vocabularies [9, 10, 11, 12, 13, 14], most of the existing electronic health data is heterogeneous among different healthcare institutions, let alone different countries. For this reason, population-based research is often slowed down by the difficulty in compiling and assessing large amounts of interoperable data. The FAIR principles [15] are a set of guidelines recently published that provide specifications on how to make data more meaningful and useful, by making it Findable, Accessible, Interoperable and Reusable. They put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting their reuse. Their adoption by data discovery platforms may facilitate data reuse and contributes to the advances of translational and clinical research.

The objective of this systematic review was to identify projects and software solutions that promote patient electronic health data discovery, as enablers for data reuse and advancement of biomedical and translational research.We have identified 20 such platforms [16, 17, 18, 19, 20, 21, 22, 23, 24, 3, 7, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34], following a semi-automatic search strategy that we present in the following section. We are interested in understanding how these systems address interoperability, how they impact clinical research and ultimately what the outcomes are in terms of scientific results and contributions to better general healthcare. Moreover, we intend

this systematic review to be a support tool for a researcher looking for EMR databases for secondary use, as we examine the existent solutions and aggregate relevant information such as type of data, number of datasets and data privacy concerns.

2. Methods

2.1. Search Strategy

Our initial approach to identifying original scientific publications showcasing biomedical data discovery platforms was to put together a list of discovery platforms with which we were familiar, given our previous experience in the area. The purpose of this was to identify common terms in the titles of the scientific work behind these platforms, so as to include them in the query of the systematic search protocol. We quickly understood this would be a very challenging task as these studies were very heterogeneous in terms of vocabulary, both in titles and associated keywords. This made it difficult to reach a common ground in what concerns the query terms. Additionally, we avoid the terms "electronic health record" or "electronic medical record" in our queries as there is a the large amount (order of thousands) of manuscripts out of scope that include these terms (such as use cases descriptions, clinical trials experiments, hardware-related). Moreover, queries such as "discovery platform" or "data discovery" on some of the most common scientific databases, such as Scopus, ACM or Pubmed, not only are not able to fully cover the chosen topic, but also lead to thousands of results related to other types of discoveries. In order to cope with these challenges, we adopted a strategy in which we limit the query terms and we explore more the notion of similarity. First, we limited the search to Pubmed database, since it indexes the most extensive collection of health-care related publications. A secondary reason for choosing Pubmed database is that it provides a public Application Programmable Interface (API) that allows programmatic retrieval of information of interest [35]. Based on the previous assessment of the lack of correlation between search concepts and scientific work, and considering the limited number of terms that can be used in querying of online scientific databases, we assumed that no single query would be able to retrieve all, or the great majority, of publications of interest. Therefore, we decided to take a cyclical machine-supported approach for identification of these publications. The programmatic retrieval was done using the Biopy-

thon framework[1]. We have limited the initial search terms to the following boolean: "data discovery[Title]" OR "discovery platform[Title]". From here, we constructed a tree search, in which the initial nodes were the Pubmed identifiers of the articles returned by the search query. For each of these ids, we automatically retrieved the first 20 most similar articles. For each of these ids, we automatically retrieved the first 20 most similar articles. We removed the duplicates by means of an automatic Python[2] script and we assessed the remaining titles for relevance for the topic. Finally, we repeated the tree search, this time looking only for the 5 most similar articles to the ones previously retrieved (excluding the initial nodes, which had already been considered by the first level of the tree search). Again, the duplicates were removed and the remaining titles were assessed. All searches were limited to the period from January 2014 to September 2018. Additionally, we discarded papers related to the area of molecular biology, since we intended to keep the focus of this paper on patient data.

After the title assessment process, we performed a manual assessment of the references cited in this initial batch of articles. The manual search complemented the initial search results by revealing a small number of manuscripts that were considered potentially relevant for the review.

2.2. Inclusion and Exclusion Criteria

The titles, list of authors and publication dates of the manuscripts resulting from the search, both systematic and manual, were joined in a list that was further ordered by author names. Multiple manuscripts belonging to the same author were analyzed in order to identify the most recent or the one that better describe the solution. Having identified one such manuscript per author, the remaining articles belonging to the same authors were discarded, as they would contain similar content and thus add some redundancy to the final results of the review. Other exclusion criteria were, as follows: duplicated entries, relevance for the chosen topic, and the level of detail to which the topic is addressed. Finally, the publication dates were reviewed in order to guarantee that only manuscripts published between January 2014 and September 2018 were included.

[1] https://biopython.org/

[2] www.python.org

2.3. Study Selection

After applying the inclusion and exclusion criteria, a total of 80 titles were considered as possible candidates for the study. These were complemented by 5 titles that were identified in their bibliography and were manually added to the review process. After abstract evaluation, 37 manuscripts were considered for full text retrieval. Of these, 20 were considered relevant for the study and will be assessed throughout the rest of this manuscript. The results and conclusions drawn are presented in the next sections, following the Preferred Reporting Items for Systematic Review and Meta-Analysis protocols (PRISMA-P) 2015 statement [36].

The complete pipeline of the methodology followed in this review is illustrated in Figure 1.

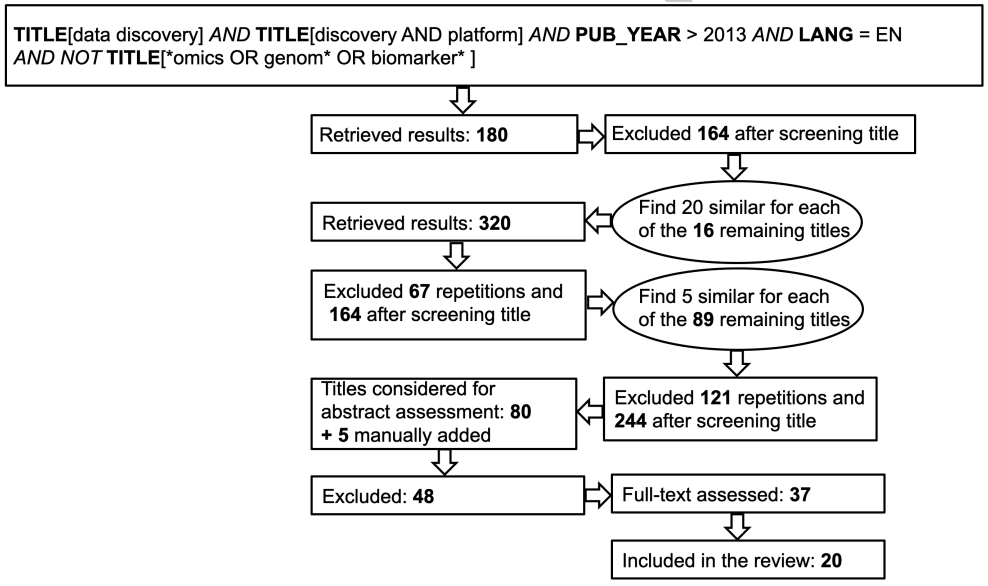


Figure 1: The systematic search process with inclusion and exclusion criteria. The ellipse shapes represent programmatic searches.

2.4. Data collection and analysis

Both authors reviewed the papers to be included in the review. They each recorded independent observations on an individual Excel spreadsheet, focusing on the type of data, data interoperability and outcomes that resulted

from use of the platforms for data discovery that were previously identified. Table 1 lists the fields that each of the authors had to track in their individual spreadsheet. Reviewers had to identify possible bias in each paper, based on the Cochrane Collaborations risk-of-bias tool [37]. Finally, observations were combined into one spreadsheet for discussion. No papers were discarded because of bias.

3. Results

We identified 20 unique publications that addressed data discovery platforms and architectures. Platforms are mainly web-sites that enable data discovery by exposing different levels of information about electronic health datasets, from meta-data to aggregated data. Architectures are fully-fledged solutions that can be installed in more than just one instance. They are usually detailed with more emphasis on technical aspects and at least one example of a platform based on such an architecture is presented.

3.1. Overview of the studies

Biomedical data exists in multiple scales, from molecular to patient data. The integration and reuse of routinely collected clinical data for research purposes has raised growing awareness within the research community. Health systems, genetics and genomics, population and public health are all areas that may benefit from big data integration and its associated technologies [38]. The secondary reuse of citizens' health data and investigation of the real evidence of therapeutics may lead to the achievement of personalized, predictive and preventive medicine [39]. However, in order for researchers to be able to reuse data and conduct integrative studies, they first have to find the right data for their research. Data discovery platforms and tools are one-stop shops that enable clinical researchers to identify datasets of interest without having to perform individual, extensive searches over distributed, heterogeneous health centers.

The manuscripts included in this review describe current data discovery solutions. While they serve the same ultimate purpose, different characteristics among these platforms can be noted. As such, we have identified platforms that integrate more specific information, for example disease-related biomedical data and platforms that showcase a broad range of life-science data. Figure 2 summarizes our findings in terms of the type of data that the identified solutions showcase.

Table 1: Characteristics that were evaluated in the review.

Field	Description
Pubmed ID	Pubmed Identifier.
Title	Title of the manuscript.
Year	Publication year.
Warehouse	Indicates whether the platform is a warehouse.
Open-access	Indicates whether the platform is open access.
Data type	What types of data does it leverage?
Data size	Number of datasets available.
FAIR	Are the FAIR guidelines addressed? If yes, how are they being followed?
API	Does the platform support interacting with the data through an API?
Ontologies/standards	Is there any standardization involved? If so, what ontologies or standards are being used?
Privacy protection	How is data privacy ensured and what are the access rules?
Relevant outcomes or use cases	What achievements has the platform enabled?
To be included	Indicates whether the reviewer considers the manuscript fit to the current review.
Comments	Any additional comments that the reviewer might have.

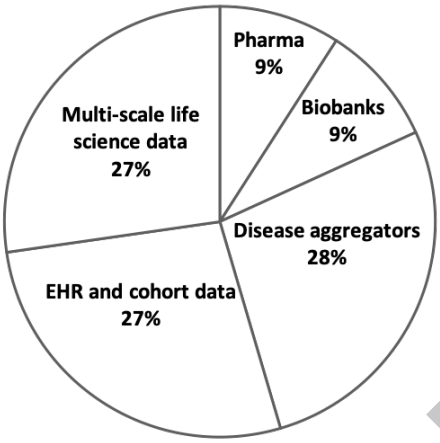


Figure 2: Scope assessment of the identified data discovery solutions.

Another relevant aspect in assessment of the identified solutions is the underlying method for exposing and eventually making data available. Discovery systems can expose data at different levels of granularity, from metadata to raw data in some cases. Data catalogs are frequent options in advertising a product and the same concept can be applied to exposing, or making biomedical data discoverable. Depending on the fundamental vision and purpose of each such system, we can distinguish two different methodologies to make data discoverable within the manuscripts selected for this review. On one hand, warehouse solutions (30%), in which all data are gathered at a single central location, and are then accessible from a single access point. On the other hand, most of the manuscripts report on discovery systems in which data remain at the owner's site, i.e. where it is collected and maintained (70%).Data never leaves the owner's site, meaning it is is kept in its original institution (owner) and no data integration is performed. In this case discovery platforms generally showcase metadata or aggregated views of the original datasets and link to the owner's site, where data can be made available following given protocols. On the other hand, data warehouses gather data from several repositories and institutions and as such the effort is duplicated as the data warehouse keeps a copy, periodically updated, of the data held at the institution level.

3.2. Summary of the studies

We now summarize the studies included in this review, taking into consideration some of the characteristics previously identified. As such, we first overview the warehouse platforms, where data can be discovered and accessed within the platform, as they are integrated at a single location. Next, we shift towards the platforms in which data is kept at the owner's site, by reviewing disease-oriented solutions. The more mature platforms are presented next, followed by more recent approaches to discoverability enabled by Linked Data. More details of each of these platforms and the type of data they promote are given in Figure 2.

Among the warehouse platforms, the Vanderbilt approach [19] is a two-in-one data warehouse, containing both fully de-identified research data and fully identified research that is made available taking into consideration access protocols and governance rules. The Project Data Sphere initiative was built to voluntarily share, integrate, and analyze historical cancer clinical trial data sets with the final goal of advancing cancer research [18]. BBMRI-ERIC, the Biobanking and BioMolecular Resources Research Infrastructure-European Research Infrastructure Consortium, is an umbrella organization for biobanking in Europe. For rare and common diseases alike, it provides fair access to quality-controlled human biological samples and associated biomedical and biomolecular data [20].

Disease oriented platforms, such as The Ontario Brain Institutes (Brain-CODE) [21] are designed with a very explicit, yet not limited, purpose of supporting researchers in better understanding a specific disease. Brain-CODE addresses the high dimensionality of clinical, neuroimaging and molecular data related with various brain conditions. The platform provides integrated datasets that can be queried and linked to provincial, national and international databases. Similarly, the breast cancer (B-CAN) platform [34] was designed as a private cancer data center that enables the discovery of cancer-related data and drives research collaborations aimed at better understanding of this disease. In the rare disease spectrum, RD-Connect [22] links genomic data with patient registries, biobanks, and clinical bioinformatics tools in an attempt to provide a FAIR rare disease complete ecosystem. In-formation coming from inventories, websites, scientific journals and technical reports can be reached from the Biobank finder, which also provides a link to the RD-Connect Sample Catalogue, an inventory of biological samples. The German Centre for Lung Research (DZL) is an association of Germany's research and medical institutions dedicated to lung research. It

provides a data warehouse where all cancer patient related data are combined and made accessible [17]. On the Alzheimer's spectrum, the Global Alzheimer's Association Interactive Network (GAAIN) is a network of shared research data, analysis tools, and computational resources to study the causes of Alzheimer's disease [28].

Among most established initiatives, Cafe Variome [23] provides a general-purpose, web-based, data discovery tool that can be quickly installed by any genotype–phenotype data owner and makes data discoverable. In this aspect, several similarities between Cafe Variome and MONTRA [24], another fully-fledged open-source discovery solution, were revealed. MONTRA is a rapid-application development framework designed to facilitate the integration and discovery of heterogeneous objects. Both solutions rely on a catalogue for data discovery, and include extensive search functionalities and query capabilities. Harvest [3] is another open-source framework of modular components, used for the rapid development and deployment of custom data discovery software applications. eGenVar [7] is a metadata cataloguing system and a software suite for reporting the presence of data from the life sciences domain. Specifically, it allows users to report, track, and share information on data content and provenance. The PopMedNet software platform features distributed querying, customizable workflows, and search capabilities [25]. A cataloguing toolkit is proposed by Maelstrom Research, built upon two main components: a metadata model and a suite of open-source software applications. When combined, the model and software support implementation of study and variable catalogues and provide a powerful search engine to facilitate data discovery. This toolkit already serves several national and international initiatives [26]. REDCap is an electronic data capture software that allows the easy building of research instruments while providing collaboration capabilities, metadata workflow, security, auditing, and export to common statistical packages. Medical librarians have used REDCap for both research data capture as well as operational databases [27]. These examples are more than just platforms, being fully-fledged architectures that can be installed in more than one site. They include software to expose biomedical datasets and different other bioinformatics tools. We consider these architectures equally relevant for the topic of this review and we assess and we assess the platforms that were built based on them.

An ambitious initiative from the US National Institute of Health (NIH), DataMed, envisions to be for data what PubMed has been for the scientific literature. Similar to the Journal Article Tag Suite used in PubMed, the

DATS model enables submission of metadata on datasets to DataMed [31]. DataMed can efficiently index and search diverse types of biomedical datasets across repositories and consists of 2 main components: the data ingestion pipeline that collects and transforms original metadata information to the DATS unified data model and a search engine that finds relevant datasets based on user-entered queries [32]. The Innovative Medicines Initiative's project EHR4CR developed a platform for reusing EHR data to support medical research. An initial instance of the platform integrated datasets from eleven participating hospitals and ten pharmaceutical companies located in seven European countries [33].

Linked Data is also explored in discovery platforms, such as Yummy-Data [29] which was designed to improve the findability and reusability of life science datasets provided as Linked Data. It consists of two components, one that periodically polls a curated list of SPARQL endpoints and a second one that monitors them and presents the information measured. BioSharing is a manually curated searchable portal of three linked registries [30] that cover standards, databases and data policies in the life sciences. Similarly, the Open PHACTS Discovery Platform [16] leverages Linked Data to provide integrated access to pharmacology databases.

3.3. Main findings

Data discovery solutions should provide intuitive, easy to use functionalities for identification of the right data. Search support is imperative for the identification of appropriate datasets of interest and for assessment of their suitability. Our findings show that all studies include this functionality. Another critical issue when dealing with biomedical data is privacy protection and proper access control; 20% of the studies take into consideration secure data access and implement access control rules, such as user permissions [24, 27, 19, 33, 21]; 15% of the platforms address the privacy issue from the point of view of data anonymization and the possibility of data owners removing their datasets from the platform [17, 40, 18]. With respect to interoperability, 40% of the evaluated solutions rely on the use of ontologies and standards so as to provide a heterogeneous view of the data [16, 17, 26, 27, 28, 30, 32, 23, 21]. This evaluation reveals a moderate adoption of the FAIR principles, which accentuate on the discovery of data by machines, through APIs (45%) [17, 26, 24, 29, 30, 32, 25, 23, 21]. Lastly, we identified a positive trend in supporting open science, as 60% of these platforms and architectures are open access [16, 26, 24, 28, 29, 30, 7, 32, 25, 3, 21, 34].

Figure 2 presents the general characteristics of the platforms and architectures. For the architectures that we have identified and that can back-up more than one platform, we include those platforms to which the original architecture publication links to. We propose this table as a discovery tool itself as it can guide researchers in finding the right platform for a specific clinical research question.

4. Discussion

Having identified the importance of data discovery platforms in the context of data interoperability and reuse, we identify current platforms and architectures that promote digital health data discovery. The majority of the solutions identified support open access and have taken the first steps in following the FAIR principles, which are recent driving forces behind establishing data interoperability. Additionally, the result table included in this review was intended as a discovery tool itself, to enable researchers to identify current platforms and assess the usability of each platform given a scientific question or a clinical research interest.

For each solutions discussed, we were interested in understanding the outcome generated, in terms of scientific and clinical contributions and the impact these platforms have on translational research. Apart from the evident contributions of some of the platforms in promoting research collaborations and shared knowledge on specific disease related research questions [28, 17, 22], some of the most mature discovery architectures, such as [27] led to more than 6000 written scientific contributions. This platform brings together 3000 institutions from 128 countries. Another relevant outcome, reached by means of the architecture presented by Davies et al. [25] led to four health data networks implemented across US: Sentinel, Patient-Centered Clinical Research Network, Massachusetts Department of Public Health, and the NIH Collaboratory Distributed Research Network. A mature pan-European discovery platform based on the architecture proposed by Silva et. al [24], the EMIF Catalogue[3] aggregates metadata of more than 10 millions European subjects in 480 datasets. Finally, the platform identified in [30] is maintained as a community resource closely embedded in and cosponsored by several infrastructure programs, including the NIH Big Data

[3] www.emif-catalogue.eu

Table 2: Summarized information about the platforms included in the review.

Citation	Open Access	Data	Ontologies/standards	Privacy Protection
Groth et al. [16] https://explorer.openphacts.org/	Yes	11 linked datasets of pharmacological and physiochemical data	Data from the various data sources are mapped into a single uniform vocabulary	Licensing terms and conditions may apply to each dataset
Majeed et al. [17] https://data.dzl.de/cometar/web/	No	17,916 patient registers of lung cancer related data; on average, 26 details are available per patient	A common terminology for patient phenotypes and bio specimen was developed and agreed upon within the consortium	Data encrypted transfer over the internet; pseudonymization on-the-fly
Bergeron et al. [26] https://www.maelstromresearch.org/maelstrom-catalogue	Yes	More than 180 cojhort epidemiological studies from 14 international networks - 6,240,000 participants	Metadata standardization	Not addressed
Silva et al. [24] https://emif-catalogue.eu/	Yes	480 datasets of cohort and EHR pan-European data	Not enforced, but possible	User permissions (Role Based Access Control) and different levels of data exposure.
Wright et al. [27] https://www.project-redcap.org/	No	Multi-scale life-science data	No direct mention. There is a CSV template for structuring the data for import (this template could lead to harmonization)	Data permissions granted through user groups and role-based access control policies
Crawford et al. [28] https://www.gaaindata.org/partners/online.html	Yes	479,483 patient records from 44 GAAIN partners containing neuroimaging, demographic, genetic, and biologic data targeting Alzheimer's disease	Partner data in GAAIN is mapped to a single schema	Data is deidentified and each custodian has a on/off data switch
Yamamoto et al. [29] http://yummydata.org/	Yes	65 Sparql endpoints amounting to 22G of multi-scale life-science Linked Data	Linked Data - Resource Description Framework (RFD) and SPARQL access protocol	Not addressed
McQuilton et al. [30] http://www.biosharing.org/	Yes	700 databases of multi-scale life-science Linked Data, including animal samples, proteins, peptides	Semantic representations of a topic or field used to catalogue and organise data into ontologies	Set of well defined policies in place for accessing data
Razick et al. [7] http://bigr.medisin.ntnu.no/data/eGenVar/	Yes	Multi-scale life-scrience data - a demo using the HUNT Biobank is in place	Metadata tags based on controlled vocabularies and standardized terms	Role based user access control policies
Chen et al. [32] https://datamed.org/	Yes	2,336,403 datasets in 75 repositories of protein, phenotype, gene expression, clinical and omics data	Yes – Data Tag Suite model (DATS)	Not addressed
Green et al. [18] https://www.projectdatasphere.org/projectdatasphere/html/	Yes	Data from 14 cancer clinical trials covering 9,000 subjects	Not addressed	Data owners are responsible for deidentifying data according to well defined standards
Danciu et al. [19] https://www.vumc.org/dbmi/synthetic-derivative	Yes	Electronic medical records of 2,2 millions uique individuals	Not addressed	User access control policies
Davies et al. [25] https://www.popmednet.org/	Yes	Multi-scale life-science data - four health data networks were implemented using PopMedNet	Yes - not specified	User access control policies
Lancaster et al. [23] https://www.cafevariome.org/	Yes	Sensitive biomedical data, from genomics to cohort data coming from 15 diagnostic or research labs	Bio-ontologies from the National Center for Biomedical Ontology (NCBO) BioPortal library	Role based discovery and access control
Gainotti et al. [22] https://platform.rd-connect.eu/	Yes	222 registries of biomolecular data and 21 biobanks	The Human Phenotype Ontology (HPO) and the Orphanet Rare Disease Ontology (ORDO)	Not addressed
Pennington et al. [3] http://harvest.research.chop.edu	No	Multi-scale life-science data from 47,300 patients across 24,900 catheterization procedures and 54,000 echocardiogram procedures within the CardioDB installation at The Children's Hospital of Philadelphia	Standard ICD9 and Current Procedural Terminology codes along with CHOP-customized codes	Data available only to CHOP partners.
De Moor et al. [33] http://www.ehr4cr.eu/views/solutions/platform.cfm	Yes	Electronic health records from 11 data provider sites in 5 European countries	EHR4CR Common Element Templates (CETs) and Common Data Elements (CDEs) can be considered as the semantic building blocks of the EHR4CR Common Information Model	The EHR4CR security architecture also mandates the adoption of a number of information security practices (e.g. regarding the treatment of passwords and personal medical data) as mandated by the EHR4CR non-functional requirements
Mayrhofer et al. [20] http://www.bbmri-eric.eu/	Yes	Human biological samples and associated biomedical and biomolecular data reaching 515 biobanks and standalone collections	Yes - not specified	Not addressed
Wen et al. [34] http://www.bcan.med.stu.edu.cn	Yes	1,086 breast cancer patients clinical data and 7 types of omics data of the Cancer Genome Atlas	No	Not addressed
Vaccarino et al. [21] https://www.axon.braininstitute.ca	No	Clinical, neuroimaging and molecular human and animal brain-related data from 40 institutions, incorporating 17,000 study participants and 1,500 animal subjects	Yes - not specified	Granular access control

to Knowledge Initiative's BioCADDIE[4] and CEDAR[5] projects. It is being enhanced as part of the Elixir UK node's contribution to the ELIXIR EXCELERATE program[6]. The amount of data collected by some platforms, such as [24, 27, 2], as well as the extent of the research collaborations that they fostered, are clear evidences that they are important enablers in reusing health data for research.

One important aspect to be considered when dealing with real world health data is data protection. Aspects such as patient privacy, informed consent, transparency or legislative frame must be considered. This responsibility lies with the data owner, who has to inform the patient and act accordingly, before making any data available for secondary use. In addition, it is essential that discovery platforms clearly define access rules and standardized permissions protocols.

These discovery platforms are quite recent enablers of translational research. They are proof that current healthcare is already impacted by the ability to find the right data source for secondary research. We believe many more positive outcomes can be expected in the near future, not only in terms of generating new knowledge but also with respect to better understanding of modern diseases and providing personalized, overall improved healthcare.

4.1. Limitations

One possible limitation of this study is the reduced number of authors that reviewed the scientific contributions included in the review. Selection bias might be present in some studies. In situations where there was no initial consent to include a manuscript, the authors presented all the reasons for the inclusion/exclusion of a given manuscript and a final decision was reached jointly. Another limitation might reside in the publication period of the considered manuscripts. An inclusion criterion was that they had to be published between January 2014 and September 2018. This decision was motivated by two factors. Firstly, we wanted to focus on recent, on-going projects that leverage biomedical data discovery. Secondly, we identified systematic reviews published previous to 2014 that partially covered this topic, such as [41].

[4]https://biocaddie.org/

[5]https://www.cedarnetwork.org.uk/

[6]https://www.elixir-europe.org/about-us/how-funded/eu-projects/excelerate

5. Conclusions

The above assessment is intended as a structured overview of the existent solutions for data discovery and integration that support data interoperability, an imminent need for biomedical data reuse. Through this study we have come to understand that data discovery platforms and tools enable more qualitative research and have the potential to speed up and reduce the cost of translational research. This review was based on the analysis of the scientific works that detail either the implementation, the use, or both aspects, of the platforms presented.

This review revealed that while there is still work to be done in this research field, the results identified are the standing proof that such platforms promote meaningful research collaboration targeted at secondary use of biomedical data.

ACKNOWLEDGEMENTS

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968 and from the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010).

References

[1] C. S. Kruse, R. Goswamy, Y. Raval, S. Marawi, Challenges and opportunities of big data in health care: a systematic review, JMIR medical informatics 4 (2016).

[2] G. Bouzillé, R. Westerlynck, G. Defossez, D. Bouslimi, S. Bayat, C. Riou, Y. Busnel, C. Le Guillou, J.-M. Cauvin, C. Jacquelinet, et al., Sharing health big data for research-a design by use cases: the inshare platform approach, in: The 16th World Congress on Medical and Health Informatics (MedInfo2017), 2017.

[3] J. W. Pennington, B. Ruth, M. J. Italia, J. Miller, S. Wrazien, J. G. Loutrel, E. B. Crenshaw, P. S. White, Harvest: an open platform for developing web-based biomedical data discovery and reporting applications, Journal of the American Medical Informatics Association 21 (2014) 379–383.

[4] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of ehr: data quality issues and informatics opportunities, Summit on Translational Bioinformatics 2010 (2010) 1.

[5] J. A. Linder, J. S. Haas, A. Iyer, M. A. Labuzetta, M. Ibara, M. Celeste, G. Getty, D. W. Bates, Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting, Pharmacoepidemiology and drug safety 19 (2010) 1211–1215.

[6] P. Huston, C. D. Naylor, Health services research: reporting on studies using secondary data sources., CMAJ: Canadian Medical Association Journal 155 (1996) 1697.

[7] S. Razick, R. Močnik, L. F. Thomas, E. Ryeng, F. Drabløs, P. Sætrom, The egenvar data management systemcataloguing and sharing sensitive data and metadata for the life sciences, Database 2014 (2014).

[8] L. Hernandez, J. Onieva, G. Fico, J. Cancela, A. Dagliati, M. Bucalo, L. Sacchi, R. Bellazzi, M. T. Arredondo, A proposal of architecture to share patients data out of healthcare settings for research purposes, in: Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on, IEEE, 2014, pp. 789–792.

[9] A. Groß, C. Pruski, E. Rahm, Evolution of biomedical ontologies and mappings: Overview of recent approaches, Computational and structural biotechnology journal 14 (2016) 333–340.

[10] S. L. Subramanian, R. R. Kitchen, R. Alexander, B. S. Carter, K.-H. Cheung, L. C. Laurent, A. Pico, L. R. Roberts, M. E. Roth, J. S. Rozowsky, et al., Integration of extracellular rna profiling data using metadata, biomedical ontologies and linked data technologies, Journal of extracellular vesicles 4 (2015) 27497.

[11] F. Shen, Y. Lee, Knowledge discovery from biomedical ontologies in cross domains, PloS one 11 (2016) e0160005.

[12] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al., The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nature biotechnology 25 (2007) 1251.

[13] B. Rance, T. Le, O. Bodenreider, Fingerprinting biomedical terminologies–automatic classification and visualization of biomedical vocabularies through umls semantic group profiles, Studies in health technology and informatics 216 (2015) 771.

[14] L. Wang, B. E. Bray, J. Shi, G. Del Fiol, P. J. Haug, A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources, Artificial intelligence in medicine 68 (2016) 47–57.

[15] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016).

[16] P. Groth, A. Loizou, A. J. Gray, C. Goble, L. Harland, S. Pettifer, Api-centric linked data integration: the open phacts discovery platform case study, Web Semantics: Science, Services and Agents on the World Wide Web 29 (2014) 12–18.

[17] R. W. Majeed, M. R. Stöhr, C. Ruppert, A. Günther, Data discovery for integration of heterogeneous medical datasets in the german center for lung research (dzl)., Studies in health technology and informatics 253 (2018) 65–69.

[18] A. K. Green, K. E. Reeder-Hayes, R. W. Corty, E. Basch, M. I. Milowsky, S. B. Dusetzina, A. V. Bennett, W. A. Wood, The project data sphere initiative: accelerating cancer research by sharing data, The oncologist 20 (2015) 464–e20.

[19] I. Danciu, J. D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, P. A. Harris, Secondary use of clinical data: the vanderbilt approach, Journal of biomedical informatics 52 (2014) 28–35.

[20] M. T. Mayrhofer, P. Holub, A. Wutte, J.-E. Litton, Bbmri-eric: The novel gateway to biobanks, Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz 59 (2016) 379–384.

[21] A. L. Vaccarino, M. Dharsee, S. C. Strother, D. Aldridge, S. R. Arnott, B. Behan, C. Dafnas, F. Dong, K. Edgecombe, R. El-Badrawi, et al.,

Brain-code: A secure neuroinformatics platform for management, federation, sharing and analysis of multi-dimensional neuroscience data, Frontiers in neuroinformatics 12 (2018) 28.

[22] S. Gainotti, P. Torreri, C. M. Wang, R. Reihs, H. Mueller, E. Heslop, M. Roos, D. M. Badowska, F. Paulis, Y. Kodra, et al., The rd-connect registry & biobank finder: a tool for sharing aggregated data and metadata among rare disease researchers, European Journal of Human Genetics 26 (2018) 631.

[23] O. Lancaster, T. Beck, D. Atlan, M. Swertz, D. Thangavelu, C. Veal, R. Dalgleish, A. J. Brookes, Cafe variome: General-purpose software for making genotype–phenotype data discoverable in restricted or open access contexts, Human mutation 36 (2015) 957–964.

[24] L. B. Silva, A. Trifan, J. L. Oliveira, Montra: An agile architecture for data publishing and discovery, Computer methods and programs in biomedicine 160 (2018) 33–42.

[25] M. Davies, K. Erickson, Z. Wyner, J. Malenfant, R. Rosen, J. Brown, Software-enabled distributed network governance: the popmednet experience, eGEMs 4 (2016).

[26] J. Bergeron, D. Doiron, Y. Marcon, V. Ferretti, I. Fortier, Fostering population-based cohort data discovery: The maelstrom research cataloguing toolkit, PloS one 13 (2018) e0200926.

[27] A. Wright, Redcap: A tool for the electronic capture of research data, Journal of Electronic Resources in Medical Libraries 13 (2016) 197–201.

[28] S. C. Neu, K. L. Crawford, A. W. Toga, Sharing data in the global alzheimer's association interactive network, Neuroimage 124 (2016) 1168–1174.

[29] Y. Yamamoto, A. Yamaguchi, A. Splendiani, Yummydata: providing high-quality open life science data, Database 2018 (2018).

[30] P. McQuilton, A. Gonzalez-Beltran, P. Rocca-Serra, M. Thurston, A. Lister, E. Maguire, S.-A. Sansone, Biosharing: curated and crowdsourced metadata standards, databases and data policies in the life sciences, Database 2016 (2016).

[31] S.-A. Sansone, A. Gonzalez-Beltran, P. Rocca-Serra, G. Alter, J. S. Grethe, H. Xu, I. M. Fore, J. Lyle, A. E. Gururaj, X. Chen, et al., Dats, the data tag suite to enable discoverability of datasets, Scientific data 4 (2017) 170059.

[32] X. Chen, A. E. Gururaj, B. Ozyurt, R. Liu, E. Soysal, T. Cohen, F. Tiryaki, Y. Li, N. Zong, M. Jiang, et al., Datamed–an open source discovery index for finding biomedical datasets, Journal of the American Medical Informatics Association 25 (2018) 300–308.

[33] G. De Moor, M. Sundgren, D. Kalra, A. Schmidt, M. Dugas, B. Claerhout, T. Karakoyun, C. Ohmann, P.-Y. Lastic, N. Ammour, et al., Using electronic health records for clinical research: the case of the ehr4cr project, Journal of biomedical informatics 53 (2015) 162–173.

[34] C.-H. Wen, S.-M. Ou, X.-B. Guo, C.-F. Liu, Y.-B. Shen, N. You, W.-H. Cai, W.-J. Shen, X.-Q. Wang, H.-Z. Tan, B-can: a resource sharing platform to improve the operation, visualization and integrated analysis of tcga breast cancer data, Oncotarget 8 (2017) 108778.

[35] E. Sayers, The e-utilities in-depth: parameters, syntax and more, Entrez Programming Utilities Help [Internet] (2009).

[36] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement, Systematic reviews 4 (2015) 1.

[37] J. P. Higgins, D. G. Altman, P. C. Gøtzsche, P. Jüni, D. Moher, A. D. Oxman, J. Savović, K. F. Schulz, L. Weeks, J. A. Sterne, The cochrane collaborations tool for assessing risk of bias in randomised trials, Bmj 343 (2011) d5928.

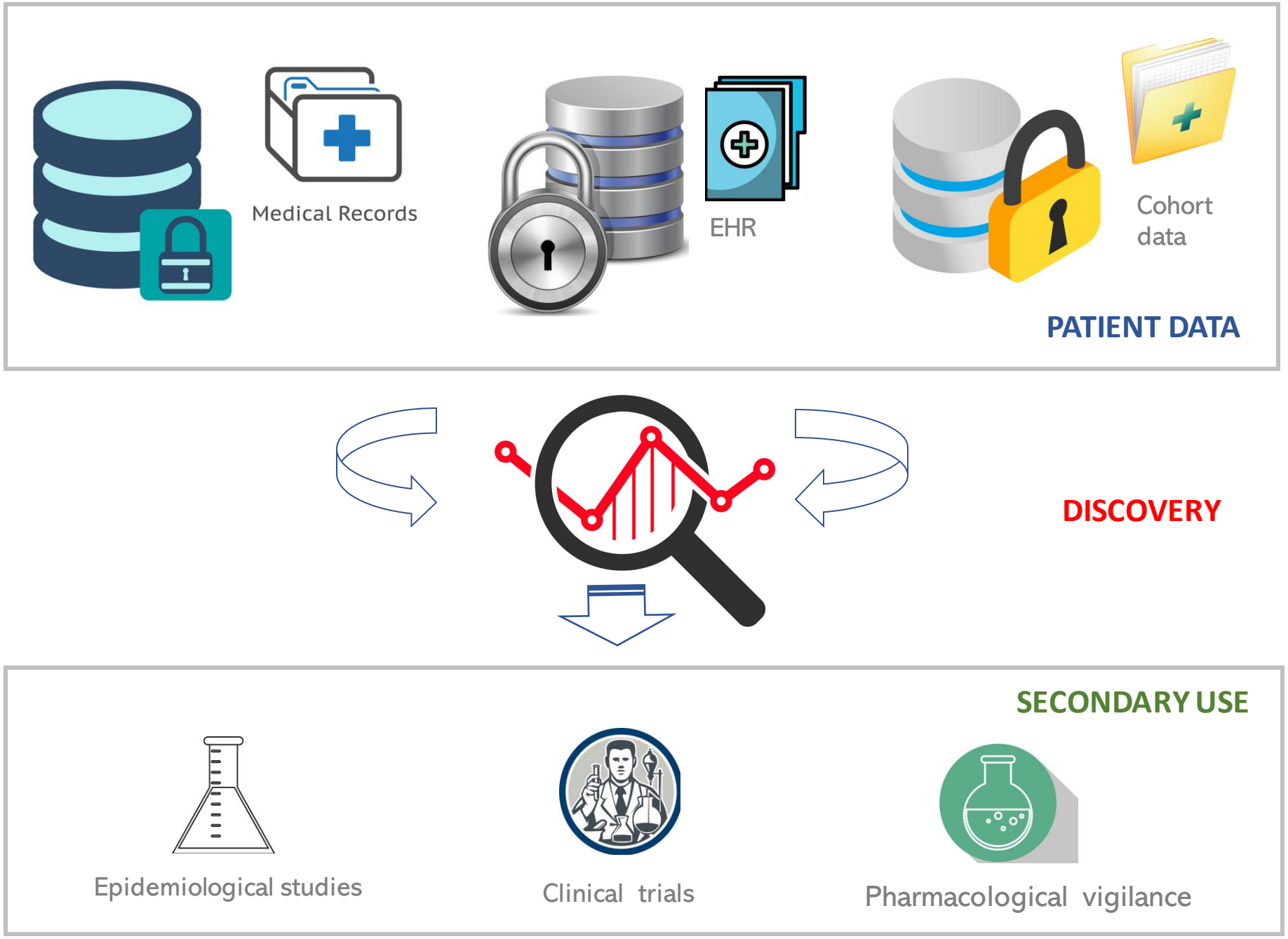
[38] F. Martin-Sanchez, K. Verspoor, Big data in medicine is driving big changes, Yearbook of medical informatics 9 (2014) 14.

[39] J. H. Phan, C. F. Quo, C. Cheng, M. D. Wang, Multiscale integration of-omic, imaging, and clinical data in biomedical informatics, IEEE reviews in biomedical engineering 5 (2012) 74–87.

[40] F. De Backere, P. Bonte, S. Verstichel, F. Ongenae, F. De Turck, Sharing health data in belgium: A home care case study using the vitalink platform, Informatics for Health and Social Care 43 (2018) 56–72.

[41] V. Canuel, B. Rance, P. Avillach, P. Degoulet, A. Burgun, Translational research platforms integrating clinical and omics data: a review of publicly available solutions, Briefings in bioinformatics 16 (2014) 280–290.

Graphical Abstract



- This systematic review identifies software platforms that enable discovery, secondary use and interoperability of electronic health data.
- A systematic search of the scientific literature published and indexed in Pubmed between January 2014 and September 2018 was performed.
- A total number of 500 candidate papers were retrieved and out of these, 85 were eligible for abstract assessment. Finally, 37studies qualified for a full-text review, and 20 provided enough information for the study objectives.
- We aggregated the identified studies in a table that was intended as a discovery tool itself, in order to enable researchers to identify current platforms and assess the usability of each platform given a scientific question or a clinical research interest.
- This review confirms that there is currently a great research interest in reusing data for secondary analysis. Data discovery platforms and tools build up more qualitative and collaborative research, as proven by many of the research collaborations that were enabled through the platforms included in this review.