Diagnosis of Rare Diseases: A scoping review of clinical decision support systems

Jannik Schaaf (jannik.schaaf@kgu.de)
Klinikum der Johann Wolfgang Goethe-Universitat Frankfurt Klinik fur Nuklearmedizin https://orcid.org/0000-0002-0058-155X
Martin Sedlmayr
Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine Technical University of Dresden
Johanna Schaefer
Medical Informatics Group (MIG), University Hospital Frankfurt
Holger Storf
Medical Informatics Group (MIG), University Hospital Frankfurt

Research article

Keywords: Rare Diseases, Computer-Assisted Diagnosis, Clinical Decision Support Systems

Posted Date: October 18th, 2019

DOI: https://doi.org/10.21203/rs.2.16205/v1

Version of Record: A version of this preprint was published at Orphanet Journal of Rare Diseases on September 24th, 2020. See the published version at https://doi.org/10.1186/s13023-020-01536-z.

Abstract

Background Rare Diseases (RD), which are defined as diseases affecting not more than 5 out of 10,000 people, are often severe, chronic, degenerative and life-threating. A main problem is the delay in diagnosis of RD. Clinical Decision Support Systems (CDSS) for RD are software-systems to support physicians in the diagnosis of patients with RD. It would therefore be useful to get a comprehensive overview of which CDSS are available and can be used under what conditions. In this work we provide a review of current CDSS in RD and which functionality and data are used by the CDSS. Methods We searched Pubmed and Cochrane for CDSS in RD published between December 1, 2008 and December 16, 2018. Only English articles, original peer reviewed journals and conference paper describing a clinical prototype or a routine use of CDSS where included. A total of 2076 articles were found and following a screening step 16 articles (describing 13 different CDSS) were considered as relevant for the final analysis. We then described and compared the CDSS using the defined categories “functionality”, “development status”, “type of clinical data” and “system availability”. Results Three types of CDSS for RD were identified: “Machine Learning and Information retrieval”, “Web Search”, and “Phenotypic and genetic matching”. 8 of the 13 reviewed CDSS are publicly available and for use by physicians. The other remaining CDSS are clinical prototypes which have been applied in clinical studies but are not accessible to others. Only one clinical prototype online. The approaches of the CDSS differ depending on what type of clinical data is used. “Machine Learning and information retrieval” can show recommendations for a diagnosis, while Web “search CDSS” will retrieve articles from literature databases (e.g. case reports), which may provide hints for a possible Diagnosis. CDSS in “Phenotypic and genetic matching” can identify similar patients based on genetic or phenotypic data. Conclusions Different CDSS for different purposes have been established and physicians have to decide which CDSS is more accurate for a particular patient case. It remains to be seen which of the CDSS will be used and maintained in the future.

Background

In the European Union (EU), a disease is declared as “rare“ if not more than 5 out of 10,000 people are affected [1]. It is estimated that about 7,000 different rare diseases (RD) exist. According to the WHO (World Health Organization), about 400 million people are affected [2]. This corresponds to approximately 6-10% of the world population. Many RD are severe, chronic, degenerative and life-threating. They also often result in impaired quality of life or severe disability [3, 4]. 80 % of RD are of genetic origin and pre-dominantly affect children, while some diseases may occur later in life [5–9].

A big challenge in the management of RD is to finding the right diagnosis for a patient. Patients with RD are sometimes diagnosed too late or not at all, especially those with phenotypes that occur later in life (e.g. Niemann-Pick disease). Patients often report many years of diagnosis odyssey [4].

In the past, several software systems have been developed to support physicians in finding the right diagnosis for patients with RD. To support the diagnostic process, so-called “Clinical Decision Support Systems“ (CDSS) can be used. We define CDSS according to Hunt et al. as follows: A system that supports clinical decision-making, in which characteristics of patients are compared with a knowledge base and the results are presented to clinicians [10]. We refer to any system following this definition “explicit CDSS”. All other systems, that a physician might use for decisions but that do not actively give recommendations based on patient characteristics, are called “implicit CDSS”.

Only two reviews of software for diagnosis support in RD are currently available. Mueller et al. [11] present an overview of software that can be used to support the diagnosis of RD. This article includes different types of software and databases that fall in both our explicit and implicit CDSS categories. In addition, only fully developed systems are presented that are available for download or online access. Systems under development, such as research prototypes or tools in clinical evaluation, are not considered. However, when developing a novel CDSS, it is necessary for developers to know which prototypes are available and which data and functions they use [10]. The second review by Svenstrup et al. gives an overview of web search, social media and data mining approaches for the diagnosis of RD. However, the focus in this article is on web search and the comparison of their own web search engine FindZebra to other search engines [12]. Despite their importance and considering the effective use of software systems for RD diagnosis, we are not aware of any reviews of developments and current systems of explicit RD CDSS. Due to the importance of improving the diagnosis of RD, highlighted by the fact that support of diagnosis of RD using software is part of national strategy plans for RD, e.g. in Germany (National Plan of Action for People with Rare Diseases [13] or the UK (The UK Strategy for Rare Diseases) [14], we address this topic in this paper.

The objective of this paper is to give a scoping review of explicit CDSS that can be used for the diagnosis support of RD. We want to give clinicians an overview of the systems that are currently available, but we also want to show researchers which functionality and data is used by the CDSS. Therefore, we define the following research questions:

(a) Which explicit CDSS are available to support diagnosis of RD patients?

(b) Which functionalities and data are used by the explicit CDSS for diagnosis support?

(c) Which of the explicit CDSS, described in the publications, can be selected by the clinicians directly and integrated into their own clinical environment?

This review is structured as follows: We will only use the term CDSS, with CDSS referring to our explicit CDSS definition, mentioned above. In the chapter “methods” we describe the data sources as well as, inclusion and exclusion criteria. We also describe how the articles were analyzed and extracted. In the “results” section, we describe findings relating to our research questions, which will be outlined and summarized in the discussion.

Methods

For bias minimization, the selection of the studies was followed by the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [15]. 21 out of 27 PRIMSA items could be considered (shown in Additional File 1).

Data Source

We retrieved articles from the Pubmed and Cochrane databases published between December 1, 2008 and December 16, 2018. The final search was conducted on December 16, 2018. We defined the following core terms for the search: "Rare Diseases" and "Clinical Decision Support". Based on this, we defined synonyms to our knowledge. The synonyms were transferred to MeSH-Terms (Medical Subject Heading) for the Pubmed and Cochrane search (shown in Table 1). The search terms were divided into four groups (Group A-D). Group "A" included two MeSH terms called "Rare Diseases" and "Orphan Diseases" as a synonym of Rare Diseases. Group "B" included several MeSH terms which were associated with Clinical Decision Support Systems based on author's experience. Group "C" was similar to Group "A" but without MeSH Terms. Group "D" included additional further terms in addition to Group "B" which were not covered by MeSH. Group "A" and "B" were combined with a logical "and", the same for group "C" and "D". This resulted in a final query (as shown in Figure 1). All search results were limited to human medicine and English language articles.

Inclusion and exclusion criteria

The inclusion criteria were defined by the type of paper and the state of the CDSS. We included English articles, original peer reviewed journals and conference paper describing a clinical prototype or a routine use of CDSS. All other publication types were excluded. We only included systems matching the definition of a CDSS mentioned in the introduction. All available titles and abstracts were screened independently by the two authors (JAS and MS). The results were entered into a spreadsheet containing the title and date of the publication, the name of the author mentioned first in the publication and the journal reference. Based on whether they corresponded to our definition of a CDSS mentioned above and if the clinical context is relating to RD, the authors (JAS and MS) decided whether to accept or reject the article. If a decision could not be made based on the abstract and title, the articles were obtained on full electronic versions via institutional library access.

Data extraction and analysis

The studies were specifically analyzed with regard to the research questions. For analysis, we defined categories based on the authors experience in RD research: "System or project name", "Functionality", "Type of clinical data", "Development status" and "System availability".

The "system or project name" is the name of the RD CDSS or the related project name. If a system or project name was unavailable, we used the name of the first author of the publication. We define "Functionality" as the technology that performs the decision support (e.g. machine learning). "Type of clinical data" indicates whether the CDSS uses clinical routine data (e.g. lab, reports, documentation) or phenotypic and genetic data. "Development status" describes if the CDSS is a clinical prototype or a fully developed system. The term "System availability" describes whether the system is available for download, whether online usage is possible or whether the system is not available.

For an unbiased overview of all implicit CDSS, the articles were grouped into the above categories. The relevant articles were entered into a table (shown in Table 2) by the authors (JAS and MS). Differences were discussed and resolved by all authors. The risk of bias for the studies was not methodically evaluated.

Results

The search identified 765 articles in Pubmed and 1311 articles in Cochrane (as shown in Figure 2). A total of 2076 articles was found, out of which 62 articles were selected following an initial screening for eligibility. After reviewing the abstract and title, 21 of the 62 articles were considered as relevant. Lastly, we checked if the articles matched our CDSS definition and excluded five articles. A total of 16 articles describing 13 different CDSS was used for data extraction and analysis. In the next section we describe the results to our selection criteria. Then the CDSS will be presented in more detail using the category "functionality".

3.1 Functionality, development status, type of clinical data, system availability

Six of the thirteen CDSS included in the analysis use phenotypic and genetic matching as a CDSS functionality [16–21] whereas three of these systems additionally make use of clinical data [18, 20, 21]. Five of the CDSS are based on machine learning and information retrieval [22–26] and two use web search methods [27, 28]. Six systems are clinical prototypes [22–24, 26, 28] and seven are full developed [16–21]. Regarding „Type of clinical data", two systems use literature databases [27, 28] and two use clinical data [24, 25], while one of these system uses both clinical and literature data [26]. Two CDSS use patient questionnaires [22, 23]. Regarding the category "System availability", eight systems are available for use via online access or download [16–21, 27] while five systems are not publicly available [22–26]. Table 3 shows a comparative overview of the mentioned results.

3.1.1 Functionality: Machine Learning and information retrieval

During a training phase, machine learning systems learn to predict or classify a problem based on existing data, such as determining whether a certain disease exists or not. After the training phase, the system can make a prediction for a new dataset [29]. Information retrieval is defined as the extraction of information in a resource to find a required piece of information. Machine learning techniques can be used in information retrieval [30].

In this review, we identified five CDSS using machine learning and information retrieval. Each article describes the CDSS as clinical prototypes which are not publicly available. All but one articles by Shen et al. [26] focus on specific RD or groups of RD [22–26], including neuromuscular diseases [22], paediatric pulmonary diseases [23], rare cancer types [25] and various rare genetic diseases [24].

Rother et al. [23] and Grigull et al. [22] used medical questionnaires completed by patients with RD to train the machine learning algorithms such as support vector machines, random forest or k-nearest neighbor. In both studies the machine learning algorithms were compared with a fusion algorithm, which is a combination of different algorithms. Both studies used patient-oriented questions to develop a questionnaire with 46 items. In the study by Grigull et al. [22] 210 patients answered the questions and a diagnosis rate of 89 % was achieved in a one-year prospective study. In the study by Rother et al. [23] the number of trained cases was slightly lower at 170, but it achieved a higher diagnosis rate with 94 % [22, 23].

In contrast to the previous studies, Sidiropoulos [25] and Garcelon [24] et al. use clinical data in their CDSS. Sidiropoulos et al. [25] developed a real time decision support for the diagnosis of rare cancer types based on histological clinical data and through on machine learning. Garcelon et al. used the Vector Space Modell (VSM) which is included in the “information retrieval" category. The authors explicitly decided against machine learning methods, since a high amount of training data would be required, which in most cases is a challenge in the field of RD [24].

Sidiropoulos et al. used quantitative histological descriptors combining structural, textural and morphological information. For real-time decision support, a GPU framework (Graphics Processing Unit) was used to show a result in real time whenever a new patient case is registered. To train the machine learning algorithm, a probabilistic neural network on 140 rare brain cancer cases was used to predict the malignancy of the tumor. Therefore, the focus was not on predicting the actual disease, as in the other articles, but on determining malignancy based on the WHO guideline for classification of neuroepithelial tumors of the central nervous system. The system achieved an accuracy of about 74 % and performed 267 to 288 times faster on the GPU-based system than on the CPU-based system (Central Processing Unit) [25].

Garcelon et al. [24] developed a system to find similar patients to an undiagnosed patient (index patient). The data is based on a clinical data warehouse containing about 400,000 patients. The similarity is calculated using the Vector Space Model (VSM), a technique of information retrieval, that computes similarity between documents represented as vectors of keywords. In this case, patient data is represented as a vector and the distance between the index patient and all other patients is calculated [24]. The evaluation of the approach was based on five different rare genetic diseases with 7 to 103 patient cases per disease. The authors evaluated the ability to find the top five patients matching the index patient as closely as possible. The percentages of index patients returning at least on true positive similar patient were reported as 94 % for Lowe Syndrome, 97 % for Epidermolysis Bulloas, 86 % for Activated PI3K Delta Syndrome, 71 % for Dowling Meara and 99 % for Rett syndrome. The processing time to retrieve similar patients in evaluation was about 12 seconds [24]. Only Sidiropoulos et al. [25] and Garcelon et al. [24] considered the processing time for their CDSS.

The approaches mentioned so far only focus on clinical data. In contrast, Shen et al. developed a system that combines clinical and literature data. The clinical data used by the authors includes 13 million unstructured clinical notes for over 700,000 patients, with the limitation to described problems and diagnosis. The literature dataset comprises about 91,000 phenotype-rare disease associations which were extracted from research articles of the SemMedDB using HPO (Human Phenotype Ontology) and GARD (Genetic and Rare Diseases Information Center) terms [26]. SemMedDb is a repository of semantic predications extracted from titles and abstracts of all Pubmed Citations, whereas GARD is a database that contains information about RD based on 4560 diseases and 32 disease categories. The HPO is the most widely used ontology describing phenotypes for genetic diseases [20]. The system developed in this study was able to combine these heterogeneous data sources into a collaborative filtering model for RD recommendation. In conclusion, the authors reported that the combination of electronic medical records and literature did not always lead to the best performance. This may be due to different approaches and expressions in medical documentation varying from physician to physician [26].

3.1.2 Functionality Web search

For complex and difficult patient cases, clinicians often consult peer-reviewed patient cases in journals, mostly case reports, to find patients with similar characteristics. This process is time-consuming and inefficient and tools to find and compare these reports would be helpful. In this review, we identified two fully developed CDSS adopting this idea of patients with RD: FindZebra [27] and a system from Taboada et al. [28].

When searching Pubmed it is often difficult to identify patients with similar characteristics. Publications with related diseases are often not marked as case reports and the number of published cases in RD is limited. Dragusin et al. [27] developed FindZebra, a search engine for RD. The authors have designed their tool similar to other search engines in order to allow an intuitively use. The search engine is based on “Indri”, an open source information retrieval system. The knowledge base of FindZebra is built on 33,144 documents of different medical cases covering 90 % of the Orphanet database. The main sources for the dataset are OMIM (Online Mendelian Inheritance in Man), GARD, Orphanet, Wikipedia and NORD (National Organization for Rare Diseases). On the other hand, Taboada et al. [28] used the Human Phenotype Ontology, the National Center of Biomedical Ontology (NCBO) and the Open Biological Ontologies (OBO). Their tool uses so called “text annotation” with the mentioned ontologies to identify phenotypes in abstracts of Pubmed. This is different from FindZebra, where the symptoms can be entered as free text. The use of the tool by Taboada et al. requires more effort by the user. In order to find corresponding phenotypes, the text must be copied into the program or read in as a file. In contrast, FindZebra provides matching results from the mentioned databases based on the symptoms entered, which are displayed in one webpage.

FindZebra uses a larger portfolio of data with 10 different data sources. While FindZebra can be accessed directly online [24], the use of the search engine by Taboada et al [28] requires a download.

3.1.3 Functionality: Genetic and phenotypic matching

A promising method when sharing patient cases is the comparison of exomes, genomes or phenotype-related patient data. Especially in RD, where 80 % of the diseases are of genetic origin, it is important to identify the external manifestation of these disorders (phenotypes) in combination with genetic testing to determine the cause of the disease (genotype) [19]. Several software systems have been developed implementing this idea. In this review, we identified the projects GeneYenta [16], GeneMatcher [17], GenIO [18], DECIPHER [19], PhenomeCentral [20] and Matchmaker Exchange [21], which we refer as “matching tools” below. In the following, we show the differences between the tools based on the listed criteria:

- (a): Available for usage: The tools are available and can be used by clinicians.
- (b): Registration necessary: A registration is required to use the platform.
- (c): Gene identification: Comparison between patients can be made on the basis of gene data.
- (d): Diagnosis code: A diagnosis code for suspected diseases can also be added for the comparison between patients.
- (e): Phenotypic Terms (HPO): Phenotypic terms are entered using the HPO nomenclature.
- (f): Acceptance of VCF (Variant Call Format) Files: Gene variants can be uploaded in the VCF format, a text file format for storing gene sequence variations.
- (g): Provides match score output: A match score for each similar patient is computed.

Table 4 shows a comparative overview of the tools with regard to the criteria above. We also highlight the aspect of data privacy, which plays a major role in the sharing of patient data.

All matching tools [16–21] are web-based and accessible online. They consider the possibility to find similar patients based on genetic or phenotypic data in the databases. A user registration is required on almost all platforms, except for GenIO [18], where an e-mail address is required to upload the data directly [18]. Furthermore, almost all matching tools support gene identification with the exception of GeneYenta [16], which focuses only on the comparison of phenotypes. To describe these phenotypes, all matching tools are using the HPO. For genetic data, PhenomeCentral [20] and GenIO [18] provide the possibility to enter these data in a VCF file format. GenIO [18], GeneMatcher [17] and PhenomeCentral [20] allow to enter suspected diseases as an additional search criteria. Each matching tool [16, 17, 19] except for GenIO [18], shows the user a match score output. For instance, GeneYenta shows a match score from 0% to 100% representing the similarity of phenotypic characteristics [16]. Since GenIO does not compare data from multiple patients, such a score is not possible. In GenIO, genetic data is entered via VCF files together with patient phenotypes using the HPO and OMIM. To process the data, GenIO uses the so-called “GenIO pipeline“, which consists of a variant annotation and phenotype processing [18]. The variant annotation uses different tools like Annovar, Anntools and SnpEff to annotate the variants of a patient. Major clinical genomic databases such as ClinVar, OMIM, the Genome Aggregation Database (gnomAD) and dbSNP (Single Nucleotide Polymorphism Database) are used as information sources for annotation [31–34]. The phenotype process of GenIO is performed with Phenolyzer which contains the list of genes related to the patient’s disease/phenotype [35]. GenIO does not consider any data sharing of patient cases, because it represents a standalone application using different clinical genomic databases [18].

Bringing the data together: The Matchmaker Exchange Project

All matching tools [16–21] are part of the Matchmaker Exchange Project (MME), which connects organizations and projects through a federate network of databases of genotypes and rare phenotypes using a common application programming interface (API) [21]. The MME enables searches across multiple databases from different platforms by making requests to all databases. To find similar matches, each request can include gene or genotype data in combination with conditions or phenotype features. MME is designed as a federated network including distributed databases which are connected through APIs to support requests. Each database can run on its own data model and its own pace [21].

Data privacy

In the following section, we describe data privacy issues and solutions for all matching tools described here. We consider how the data is shared with third parties and how the access to the data is managed. PhenomeCentral, DECIPHER and Matchmaker Exchange include concepts for the data visibility based on different levels [19–21]. For instance, patient records in PhenomeCentral have different visibility settings. These levels are “private”, “matchable”, and “public”. If “private” is selected, data is only visible for the submitter and is not available for matchmaking. For “matchable”, the submitter can see other similar patients and other users can retrieve this patient’s data, i.e. the own dataset is used for matching. However, Genomic and phenotype information is not visible. More precisely, phenotypes become generalized and are shown on gene-level only. The third level, “public”, is more open. The patient record is visible to all registered users and available for matching. Similar patients are shown to the submitters and phenotypes and genomic variants are visible [20].

GeneYenta has less security requirements. It only allows to store HPO-based phenotype data and does not include any patient-identifying data. The idea is that data sharing is performed outside the platform and according to the rules of the respective institutions [16, 20]. This concept is similar to

GeneMatcher, where no identifiable data of the patient is provided to other clinicians. Data submitters have full control over their data (e.g. gene name, phenotypic features) and can delete or edit it at any time. Users only see their own data. Further level for data sharing are not described [17].

Discussion

Within the last ten years, several CDSS to support diagnosis of RD have been developed. We performed a literature research to conduct a scoping review of CDSS in RD. The search was limited to articles published between December 1, 2008 and December 16, 2018. Since the number of CDSS in RD is rather low compared to other types of CDSS (e.g. cardiovascular diseases 45 [36], medication safety 237 [37], antimicrobial management 38 [38]), we chose a period of ten years for the search.

In this review 13 different CDSS described in 16 articles were identified. 8 of the 13 CDSS are publicly available and can be used by clinicians.

We identified three categories which are used in CDSS for RD: “Machine learning and information retrieval“, “web search“ and “genetic and phenotypic matching“. These approaches mainly differ in what type of data is used for decision support. For an undiagnosed patient, machine learning and information retrieval algorithms show the clinician a recommendation for a diagnosis.

“Web search CDSS” provide the best fitting literature data to a query and genetic and phenotypic portals provide the best match of similar patients. The studies about “machine learning” identified in this review mainly use different types of clinical data (e.g. clinical notes, histological data or questionnaires) stored in electronic health records or other systems in the clinical environment. Only Shen et al. [26] covered the combination of literature data and clinical data via data-fusion strategies. The web search CDSS identify similar patient cases and were introduced by Dragusin et al. [27] and Taboada et al. [28]. Both approaches use data from literature and known databases of rare and genetic diseases to identify relevant cases. The third category "genetic and phenotypic matching“ uses genetic and phenotypic data to identify similar patients.

Looking more closely at the use of data in the “machine learning and information retrieval” category, all studies consider only one RD or a limited set of RD. The application of these tools to a larger patient cohort with heterogeneous data is therefore only possible if sufficient patient data is available for the corresponding cohort. Due to this problem, Garcelon et al. [24] explicitly do not use machine learning algorithms and focus on an approach using the VSM. However, the authors of the study do not compare their approach with machine learning algorithms, hence, no statement can be made about performance differences between machine learning and VSM.

Both “web search“ and “genotypic and phenotypic matching“ have the advantage of a larger knowledge base, so search queries are not limited to a certain disease or disease group. However, there is a need for further studies, e.g. for FindZebra, which should do further tests even with a large number of heterogeneous queries. Patient characteristics can be registered directly and a result is returned immediately. The system is permanently available and, due to its simplicity, can be used in everyday clinical practice. A limitation is that results cannot be exported and the system cannot be adapted to individual needs of patient data. For instance, no upload of patient data from medical records or existing systems is possible.

Regarding to genotypic and phenotypic matching, the idea is to share patient cases of undiagnosed patients with the goal to identify similar patients across the world. It brings another opportunity than searching for patient characteristics in literature or discuss them on conferences. But describing phenotypes is a challenging task. Individual patients are often not completely described, for instance, patient is not reporting all current symptoms of his disease. If an anomaly in an individual patient is not described it does not mean that this anomaly does not exist for a particular disease. The description of the phenotype features also depends on the clinician’s experience. Another problem is that one phenotype can be caused by multiple genetic defects. Therefore, the authors claim that they do not only use phenotypes for matching [18].

A limitation of this review is that there may be other software systems that can support the diagnosis of RD, but either do not match with definition of a CDSS or were not found by the search query. One system we know from our own experience is Phenomizer [11]. This tool was not found by the search query, but is routinely used in clinical environments. Phenomizer facilitates the work of differential diagnosis using the HPO for entering phenotypes. The software classifies all diseases listed in OMIM, Orphanet and DECIPHER according to a score and calculates how the phenotypic characteristics match the diseases [39]. Another tool is Isabel Healthcare, which is recommended by Mueller et al. [11]. This is a web-based diagnosis support system, which has not been developed specifically for RD, but covers a broad medical context. The database of the system includes 11,000 diagnoses and 4,000 pharmaceuticals and is updated every three months. Isabel Healthcare consists of two components: A diagnosis checklist and a knowledge component. The diagnosis checklist creates a weighted list of differential diagnoses based on entered symptoms. For each differential diagnosis, a search in medical databases like Pubmed, CaseReports or Up2Date can be performed [11].

A further limitation of this review is that no data was collected relating to the effectivity or usability of CDSS in the clinical settings. This topic was not covered by the individual articles reviewed here. All articles evaluated the performance of their respective methods but did not describe the user experience in clinical application. We consider it as necessary to involve users in the development of CDSS, especially for clinical prototypes [29].

Conclusions

In summary, this review shows that some software-systems are under development and some are readily available to support the diagnoses of RD. A clinician has to decide which system can be used for which purpose and at what stage of the diagnosis process, based on his experience and the

respective patient case. Looking ahead, it remains interesting which of the mentioned tools in this review will be further developed and actively used. This could be a subject of another study in the future.

In the case of clinical prototypes, the current literature does not allow a statement whether these tools will be further developed in the future. Only Grigull et al. [22] indicates that their tools are limited to the known diseases and categories and therefore not ready for clinical use. Looking at the published systems and their usage statistics (July 2019), DECIPHER involves 3000 patients, is cited by 2000 publications and was last updated in May 2019 [40]. GeneMatcher has 33124 submissions but no information on the last software update was found [41]. Matchmaker contains about 68165 cases and no information of the current software version is available [42]. PhenomeCentral includes 7368 cases and is currently available in version 1.1.16 [42]. For Taboada et al., GeneYenta and GenIO no statistics were available. GenIO was last updated in 2017 and seems to be inactive [43]. For FindZebra no usage statistic is available, but we know from our own experience, that this tool is used by some German experts for RD.

Abbreviations

API: Application Programming Interface; CDSS: Clinical Decision Support Systems; CPU: Central Processing Unit; dbSNP: Single Nucleotide Polymorphism Database; European Union: EU; GARD: Genetic and Rare Diseases Information Center; gnomAD: Genome Aggregation Database; GPU: Graphics Processing Unit; HPO: Human Phenotype Ontology; MeSH: Medical Subject Heading; MME: Matchmaker Exchange Project; NCBO: National Center of Biomedical; NORD: National Organization For Rare Diseases; OBO: Open Biological Ontologies; OMIM: Online Mendelian Inheritance in Man; RD: Rare Diseases; VCF: Variant Call Format; Vector Space Model: VSM; World Health Organization: WHO

Declarations

Author's contributions

JAS and MS designed the review, formulated the research questions, defined inclusion and exclusion criteria and conducted the Pubmed and Cochrane search. JAS and MS also performed the data extraction and analysis. HS and JOS advised the process by reviewing and revising the results of JAS and MS. They also contributed to data analysis by reviewing the results of JAS and MS. The first draft of the article was written by JAS, whereas MS reviewed this process. HS and JOS revised the article and provided valuable input and comments. The final manuscript was written by JS and approved by all authors.

Acknowledgements

This study was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. med" from the Carl Gustav Carus Faculty of Medicine Technical University of Dresden (J Schaaf).

Funding

No funding was obtained for this study.

Availability of data and materials

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Authors details

[1] Medical Informatics Group (MIG), University Hospital Frankfurt, Frankfurt, Germany

[2] Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine Technische Universität Dresden, Dresden, Germany

References

1. Nabarette H, Oziel D, Urbero B, Maxime N, Ayme S. [Use of a directory of specialized services and guidance in the healthcare system: the example of the Orphanet database for rare diseases]. Rev Epidemiol Sante Publique. 2006;54:41–53.

2. World Health Organization. Priority diseases and reasons for inclusion. 2013. https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf.
3. Taruscio D, Floridia G, Salvatore M, Groft SC, Gahl WA. Undiagnosed Diseases: Italy-US Collaboration and International Efforts to Tackle Rare and Common Diseases Lacking a Diagnosis. Adv Exp Med Biol. 2017;1031:25–38.
4. Evans WR, Rafi I. Rare diseases in general practice: recognising the zebras among the horses. Br J Gen Pract. 2016;66:550–1.
5. Guillem P, Cans C, Robert-Gnansia E, Aymé S, Jouk P. Rare diseases in disabled children: an epidemiological survey. Arch Dis Child. 2008;2:115–8.
6. Zurynski Y, Frith K, Leonard K, Elliot E. Rare childhood diseases: how should we respond? Arch Dis Child. 2008;93:1071–4.
7. Denis A, Mergaert L, Fostier C, Cleemput I, Simoens C. A comparative study of European rare disease and orphan drug markets. Health Policy. 2010;97:173–9.
8. Griffon N, Schuers M, Dhombres F, Merabti T, Kerdelhue G, Rollin L, et al. Searching for rare diseases in PubMed: a blind comparison of Orphanet expert query and query based on terminological knowledge. BMC Med Inform Decis Mak. 2016;16:101.
9. Rare Disease UK. What is a Rare Diseases. 2018. https://www.raredisease.org.uk/what-is-a-rare-disease/. Accessed 20 Sep 2019.
10. Hunt D, Haynes R, Hanna S, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA. 1998;280:1339–1346.
11. Mueller T, Jerrentrupp A, Schäfer J. Computerunterstützte Diagnosefindung bei seltenen Erkrankungen. Internist. 2018;59:391–400.
12. Svenstrup D, Jorgensen HL, Winther O. Rare disease diagnosis: A review of web search, social media and large-scale data-mining approaches. Rare Dis Austin Tex. 2015;3:e1083145.
13. Geschäftsstelle des Nationalen Aktionsbündnisses für Menschen mit Seltenen Erkrankungen (NAMSE). National action league for people with rare diseases. 2010. http://www.namse.de/images/stories/Dokumente/Aktionsplan/national%20plan%20of%20action.pdf. Accessed 11 Jan 2019.
14. Departement of Health. The UK Strategy for Rare Diseases. 2013. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/260562/UK_Strategy_for_Rare_Diseases.pdf. Accessed 11 Sep 2019.
15. Liberati A, Altman D, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS medicine. 2009;6:e1000100.
16. Gottlieb MM, Arenillas DJ, Maithripala S, Maurer ZD, Tarailo Graovac M, Armstrong L, et al. GeneYenta: a phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation. Hum Mutat. 2015;36:432–8.
17. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat. 2015;36:928–30.
18. Koile D, Cordoba M, Sousa Serro M, Kauffman MA, Yankilevich P. GenIO: a phenotype-genotype analysis web server for clinical genomics of rare diseases. BMC Bioinformatics. 2018;19:25.
19. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. 2014;42 Database issue:D993–1000.
20. Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, et al. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum Mutat. 2015;36:931–40.
21. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat. 2015;36:915–21.
22. Grigull L, Lechner W, Petri S, Kollewe K, Dengler R, Mehmecke S, et al. Diagnostic support for selected neuromuscular diseases using answer-pattern recognition and data mining techniques: a proof of concept multicenter prospective trial. BMC Med Inform Decis Mak. 2016;16:31.
23. Rother A-K, Schwerk N, Brinkmann F, Klawonn F, Lechner W, Grigull L. Diagnostic Support for Selected Paediatric Pulmonary Diseases Using Answer-Pattern Recognition in Questionnaires Based on Combined Data Mining Applications–A Monocentric Observational Pilot Study. PloS One. 2015;10:e0135180.
24. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform. 2017;73:51–61.
25. Sidiropoulos K, Glotsos D, Kostopoulos S, Ravazoula P, Kalatzis I, Cavouras D, et al. Real time decision support system for diagnosis of rare cancers, trained in parallel, on a graphics processing unit. Comput Biol Med. 2012;42:376–86.
26. Shen F, Liu S, Wang Y, Wen A, Wang L, Liu H. Utilization of Electronic Medical Records and Biomedical Literature to Support the Diagnosis of Rare Diseases Using Data Fusion and Collaborative Filtering Approaches. JMIR Med Inform. 2018;6:e11301.
27. Dragusin R, Petu C, Lioma C, Jorgensen H, Cox I, Hansen L, et al. FindZebra: a search engine for rare diseases. Int J Med Inform. 2013;82:528–538.
28. Taboada M, Rodriguez H, Martinez D, Pardo M, Sobrido MJ. Automated semantic annotation of rare disease cases: a case study. Database J Biol Databases Curation. 2014;2014:1–13.
29. Fraccaro P, O'Sullivan D, Plastiras P, O'Sullivan H, Dentone C, Di Biagio A, et al. Behind the screens: Clinical decision support methodologies - A Review. Health Policy and Technology. 2015;4:29–38.
30. Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: Addison-Wesley; 1999.

31. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
32. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
33. Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinforma Oxf Engl. 2012;28:724–5.
34. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
35. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12:841–3.
36. Njie GJ, Proia KK, Thota AB, Finnie RKC, Hopkins DP, Banks SM, et al. Clinical Decision Support Systems and Prevention: A Community Guide Cardiovascular Disease Systematic Review. Am J Prev Med. 2015;49:784–95.
37. Jia P, Zhang L, Chen J, Zhao P, Zhang M. The Effects of Clinical Decision Support Systems on Medication Safety: An Overview. PLOS ONE. 2016;11:e0167683.
38. Rawson TM, Moore LSP, Hernandez B, Charani E, Castro-Sanchez E, Herrero P, et al. A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately? Clin Microbiol Infect. 2017;23:524–32.
39. Köhler S, Schulz M, Krawitz P, Bauer S, Dölken S, Mundlos C, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85:457–64.
40. DECIPHER. Happy 15th Birthday DECIPHER. https://decipher.sanger.ac.uk/. Accessed 15 Sep 2019.
41. Sobreira N, Schiettecatte F. GeneMatcher - Statistics. 2019. https://www.genematcher.org/statistics/. Accessed 4 Sep 2019.
42. Philippakis A. Matchmaker Exchange Statistics and Publications. 2019. https://www.matchmakerexchange.org/statistics.html. Accessed 4 Sep 2019.
43. Koile D. Clinical Genomics Assistant Tool. 2017. https://bioinformatics.ibioba-mpsp-conicet.gov.ar/GenIO/index.php. Accessed 4 Sep 2019.

Tables

Table 1: Search terms of Pubmed and Cochrane search

Group A	Group B	Group C	Group
Rare Diseases	Medical Informatics Applications	rare diseases	Search engines
Orphan Diseases	Algorithms	rare diseases	Diagnostic decision support
	Diagnostic Screening Programs	orphan diseases	
	Diagnostic services		
	Computer-assisted Diagnosis		
	Clinical Prediction Rule		
	Decision Aids		
	Decision Support Technics		
	Models, Decision Support		
	Clinical Decision Support		
	Decision Support, Clinical		
	Analysis, Decision		
	Decision, Analysis		
	Decision Modeling		
	Clinical Decision Support Systems		
	Computer-Assisted Decision Making		
	Medical Decision Making, Computer Assisted		
	Computer-Assisted Diagnosis		
	Data Warehousing		
	Medical Informatics		
	Software		

Legend: Search for Group A and B where performed with MeSH-Terms and Group C and D without MeSH-Terms

Table 2: Final relevant articles

No.	Paper Title	System or project name	Functionality	System availability	Type of clinical data	Development status
1	Automated semantic annotation of rare disease cases: a case study	Taboada et al.	Web search	Download available via URL: http://www.usc.es/keam/PhenotypeAnnotation/	Literature database	Clinical prototype
2	DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation	DECIPHER	Phenotypic and genetic matching	Online usage possible via URL: https://decipher.sanger.ac.uk/	Phenotypic and genetic data	Full developed system
3	Diagnostic support for selected neuromuscular diseases using answer-pattern recognition and data mining techniques: a proof of concept multicenter prospective trial	Grigull et al.	Machine Learning	System is not available	Patients questionnaire	Clinical Prototype
4	Diagnostic Support for Selected Paediatric Pulmonary Diseases Using Answer-Pattern Recognition in Questionnaires Based on Combined Data Mining Applications--A Monocentric Observational Pilot Study	Rother et al.	Machine Learning	System is not available	Patients questionnaire	Clinical Prototype
5	Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack	Garcelon et al.	Information Retrieval	System is not available	Clinical data	Clinical Prototype
6	FindZebra: a search engine for rare diseases	FindZebra	Web search	Online usage possible via URL: http://www.findzebra.com/	Literature database	Full developed system
7	GeneMatcher: a matching tool for connecting investigators with an interest in the same gene	GeneMatcher	Phenotypic and genetic matching	Online usage possible via URL: https://genematcher.org/	Phenotypic and genetic data	Full developed system
8	GeneYenta: a phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation.	GeneYenta	Phenotypic and genetic matching	Online usage possible via URL: https://geneyenta.com	Phenotypic Data	Full developed system
9	GenIO: a phenotype-genotype analysis web server for clinical genomics of rare diseases	GenIO	Phenotypic and genetic matching	Online usage possible via URL: https://bioinformatics.ibioba-mpsp-conicet.gov.ar/GenIO/index.php	Genetic, phenotypic and clinical data	Full developed system
10	Matchmaker Exchange	Matchmaker	Phenotypic and genetic matching	Online usage possible via URL: https://www.matchmakerexchange.org/	Genetic, Phenotype and clinical data	Full Developed system
11	Real time decision support system for diagnosis of rare cancers, trained in parallel, on a graphics processing unit	Sidiropoulos et al.	Machine Learning	System is not available	Clinical data	Clinical Prototype
12	Specialized tools are needed when searching the web for rare disease diagnoses	FindZebra	See No 6.	See No 6.	See No 6.	See No 6.
13	The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles	Matchmaker Exchange	See No 10.	See No 10.	See No 10.	See No 10.
14	The Matchmaker Exchange: a platform for rare disease gene discovery	Siehe Matchmaker	See No 10.	See No 10.	See No 10.	See No 10.

		Exhange				
15	Utilization of Electronic Medical Records and Biomedical Literature to Support the Diagnosis of Rare Diseases Using Data Fusion and Collaborative Filtering Approaches	Shen et al.	Information Retrieval	System is not available	Literature Database and clinical data	Clinical Prototype
16	PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases	Buske et al.	Phenotypic and genetic matching	Online usage possible via URL: https://www.phenomecentral.org/	Genetic, Phenotype and clinical data	Full Developed System

Table 3: Results of the categories "Functionality", "Type of clinical data", "Clinical prototypes" and "System availability"

Category	Subcategories	Total/ Frequency
Functionality	Phenotypic and genetic matching	6 (46.15 %)
	Machine learning and Information retrieval	5 (38.46 %)
	Web search	2 (15.38%)
Type of clinical data	Phenotypic and genetic data	6 (46.15 %)
	Literature databases	3 (23.07 %)
	Patient questionnaires	2 (15.38 %)
	Clinical data	2 (15.38 %)
Clinical Prototypes	Full developed systems	7 (53.84 %)
	Clinical prototypes	6 (46.15 %)
System availability	Online access or download available	8 (61.35 %)
	Not publicly available	5 (38.46 %)

Table 4: Overview of CDSS in Genetic and phenotypic matching

	Available for usage	Registration necessary	Gene identification	Diagnosis code	Phenotypic terms (HPO)	Acceptance of VCF files	Provides match score output
DECIPHER	Yes	Yes	Yes	No	Yes	No	Yes
GeneYenta	Yes	Yes	No	No	Yes	No	Yes
GenIO	Yes	No	Yes	Yes	Yes	Yes	No
GeneMatcher	Yes	Yes	Yes	Yes	Yes	No	Yes
PhenomeCentral	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Figures

(Rare Diseases[MeSH Terms] AND Medical Informatics Applications[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Algorithms[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Diagnostic Screening Programs[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Diagnostic Services[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Computer-Assisted Diagnosis[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Clinical Prediction Rule[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Decision Aids[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Decision Support Technics[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Models, Decision Support[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Clinical Decision Support[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Decision Support, Clinical[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Analysis, Decision[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Decision Analysis[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Decision Modeling[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Clinical Decision Support Systems[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Computer-Assisted Decision Making[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Medical Decision Making, Computer-Assisted[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Computer-Assisted Diagnosis[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Data Warehousing[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Medical Informatics[MeSH Terms])
OR (Rare Diseases[MeSH Terms] AND Software[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Computer-Assisted Diagnosis[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Algorithms[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Diagnostic Screening Programs[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Diagnostic Services[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Computer-Assisted Diagnosis[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Clinical Prediction Rule[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Decision Aids[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Decision Support Technics[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Models, Decision Support[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Clinical Decision Support[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Decision Support, Clinical[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Analysis, Decision[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Decision Analysis[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Decision Modeling[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Clinical Decision Support Systems[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Computer-Assisted Decision Making[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Medical Decision Making, Computer-Assisted[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Computer-Assisted Diagnosis[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Data Warehousing[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Medical Informatics[MeSH Terms])
OR (Orphan Diseases[MeSH Terms] AND Software[MeSH Terms])
OR ((rare diseases) AND search engines)
OR ((rare diseases) AND search engine)
OR ((rare diseases) AND diagnostic decision support)
OR ((rare disease) AND search engines)
OR ((rare disease) AND search engine)
OR ((rare disease) AND diagnostic decision support)
OR ((orphan diseases) AND search engines)
OR ((orphan diseases) AND search engine)
OR ((orphan diseases) AND diagnostic decision support)
OR ((orphan disease) AND search engines)
OR ((orphan disease) AND search engine)
OR ((orphan disease) AND diagnostic decision support)

Figure 1

Search Query

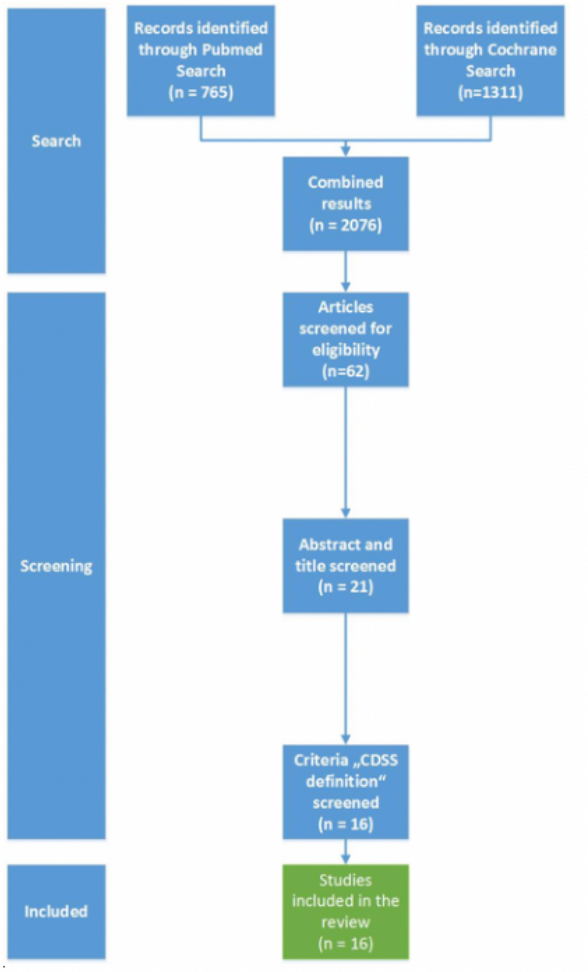

Figure 2

Search process of article selection

Supplementary Files

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

- PRISMA2009checklist.doc