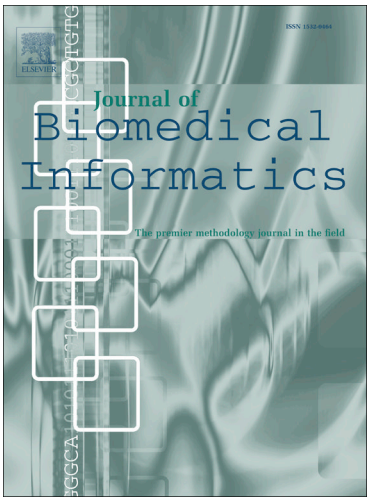

Patient similarity for precision medicine: a systematic review

E. Parimbelli, S. Marini, L. Sacchi, R. Bellazzi

PII: S1532-0464(18)30107-2
DOI: https://doi.org/10.1016/j.jbi.2018.06.001
Reference: YJBIN 2990

To appear in: *Journal of Biomedical Informatics*

Received Date: 31 January 2018
Revised Date: 16 May 2018
Accepted Date: 1 June 2018

Please cite this article as: Parimbelli, E., Marini, S., Sacchi, L., Bellazzi, R., Patient similarity for precision medicine: a systematic review, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.06.001

1)Patient similarity for precision medicine: a systematic review

E. Parimbelli[ad] *, S. Marini[bd], L. Sacchi[cd], R. Bellazzi[cde]

[a] *Telfer school of Management, University of Ottawa, Ottawa, Canada*
[b] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Harbor, USA*
[c] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy*
[d] *Interdepartmental Centre for Health Technologies, University of Pavia, Italy*
[e] *IRCCS ICS Maugeri, Pavia, Italy*

Corresponding author:
Enea Parimbelli
Telfer school of Management
University of Ottawa
55 Laurier E., Ottawa, ON, K1N 6N5 Canada
+1 613 562 5800 ext. 4322 (office)
enea.parimbelli@gmail.com

Patient similarity for precision medicine: a systematic review

E. Parimbelli[ad] *, S. Marini[bd], L. Sacchi[cd], R. Bellazzi[cde]

[a] *Telfer school of Management, University of Ottawa, Ottawa, Canada*
[b] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Harbor, USA*
[c] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy*
[d] *Interdepartmental Centre for Health Technologies, University of Pavia, Italy*
[e] *IRCCS ICS Maugeri, Pavia, Italy*

Abstract

Evidence-based medicine is the most prevalent paradigm adopted by physicians. Clinical practice guidelines typically define a set of recommendations together with eligibility criteria that restrict their applicability to a specific group of patients. The ever-growing size and availability of health-related data is currently challenging the broad definitions of guideline-defined patient groups. Precision medicine leverages on genetic, phenotypic, or psychosocial characteristics to provide precise identification of patient subsets for treatment targeting. Defining a patient similarity measure is thus an essential step to allow stratification of patients into clinically-meaningful subgroups. The present review investigates the use of patient similarity as a tool to enable precision medicine. 279 articles were analyzed along four dimensions: data types considered, clinical domains of application, data analysis methods, and translational stage of findings. Cancer-related research employing molecular profiling and standard data analysis techniques such as clustering constitute the majority of the retrieved studies. Chronic and psychiatric diseases follow as the second most represented clinical domains. Interestingly, almost one quarter of the studies analyzed presented a novel methodology, with the most advanced employing data integration strategies and being portable to different clinical domains. Integration of such techniques into decision support systems constitutes and interesting trend for future research.

Keywords: patient similarity, precision medicine, patient subgroups, data integration

1 Introduction

Since the introduction of evidence-based medicine in the early 1990s, the way physicians manage their patients has been strongly driven by clinical practice guidelines. A clinical practice guideline is a collectively produced document, informed by a systematic review of evidence, that groups the most updated available knowledge for the treatment of a disease with the purpose of disseminating such

knowledge and standardize care to ensure highest quality[1]. A guideline typically defines a set of recommendations together with eligibility criteria that restrict their applicability to a specific group of patients. In presence of a new case, physicians typically select the most appropriate recommendations comparing the characteristics of the patient under evaluation to the ones of the guideline-defined subgroups, and plan treatment accordingly. Adherence to a guideline is an important factor defining quality of care, but there are situations where a deviation is desirable and helps addressing the needs and peculiarities of those cases. More specifically, the ever-growing size and availability of health-related data is currently challenging the broad definitions of patients groups that most clinical guideline recommendations provide, demanding a more precise identification of targets for treatments.

In recent years, research efforts in this direction have been labelled with the term *precision medicine*, especially after former president Obama's launch of the precision medicine initiative in early 2015[2]. The primary aim of this line of research is to improve clinical outcomes for individual patients through more precise treatment targeting by leveraging on genetic, biomarker, phenotypic, or psychosocial characteristics that distinguish a given patient from others with similar clinical presentations[3]. However, the promise of precision medicine to provide highly personalized and highly effective treatment needs to face a substantial challenge: the more data we collect about a single patient, the more we are able to tell he/she is different from any other, making each case *unique*. This fact is potentially disruptive for the entire evidence-based medicine paradigm which, on the other hand, relies on the possibility to group patients in significant subsets to enable the careful design of clinical studies, sound statistical testing of medical hypotheses, and the possibility to generalize findings of research conducted on a limited set of individuals to a wider population of patients. Defining the proper granularity for such patient subsets is a challenging task. A tradeoff between evidence extracted from large groups of patients to reach significance and sensitivity to latest precision-medicine research findings, which may lead to treat each patient as a unique sample, needs to be reached. Similarity between patients is one of the most promising tools to address this challenge. Defining a similarity measure that is able to deal with the high-dimensional space of patient data is an essential step to allow stratification of patients into clinically-meaningful subgroups.

In this paper we present the results of a systematic review of the literature we conducted to investigate the use of patient similarity as a tool to enable precision medicine. Our primary objectives are to identify what types of data are employed in defining patient similarity in different clinical domains, what are the main methodologies used to calculate such a similarity, and what are the results that studies employing patient similarity have been able to obtain.

2 Methods

2.1 Search strategy

We conducted a literature search of the PubMed, Scopus and IEEEXplore repositories for studies published in the last 6 years (2012- 2017) employing patient similarity for precision medicine purposes. The last search was conducted on October 17, 2017. Our search strategy included a combination of terms related to patient similarity (in a disjunction) in conjunction with a set of terms aiming to restrict the retrieved studies to the precision medicine context (also in a disjunction). The complete search strings employed in the three databases are reported in supplementary table 2. Duplicate results were removed. Titles and abstracts of each paper were screened, and irrelevant articles removed before full-text analysis.

2.2 Eligibility criteria

We considered articles written in English and published in peer-reviewed journals from January 2012 to October 2017. After initial screening of title and abstract we excluded articles meeting any of the following criteria:

1. The study focuses on non-human subjects
2. The study describes the development of a new laboratory procedure or other instrument (e.g. questionnaire)
3. The study is a literature review or meta-analysis or position paper
4. The article is not about similarity between patients
5. The article describes a clinical trial design and/or execution
6. The study consists in statistical hypotheses testing
7. The study only attempts to validate previously known patient subgroups

8. The article is a single case report

Note that review articles were excluded from the review itself, but were used to inform the discussion in the present article. Similarly, the references of such review papers were screened to identify relevant articles which, when found, where added to the list of articles considered in the present review. The complete list of articles analyzed in this review is published with the article (and its Supplementary Information files).

3 Results

We initially retrieved 782 articles using the search strategy specified in the Methods section. The additional sources (e.g. screening of references of retrieved review articles) allowed us to add further 14 articles, which we comment in the discussion section. 97 papers were identified as duplicates and removed. An initial screening based on title and abstract was performed and articles not matching the previously defined eligibility criteria were excluded. Full text articles were then accessed for deeper eligibility checking, analyses and categorization. The final list of studies included in the review consists in 279 articles (see supplementary table 1). Figure 1, following the reporting standards defined by the PRISMA statement (http://www.prisma-statement.org), summarizes the article selection process and its results.

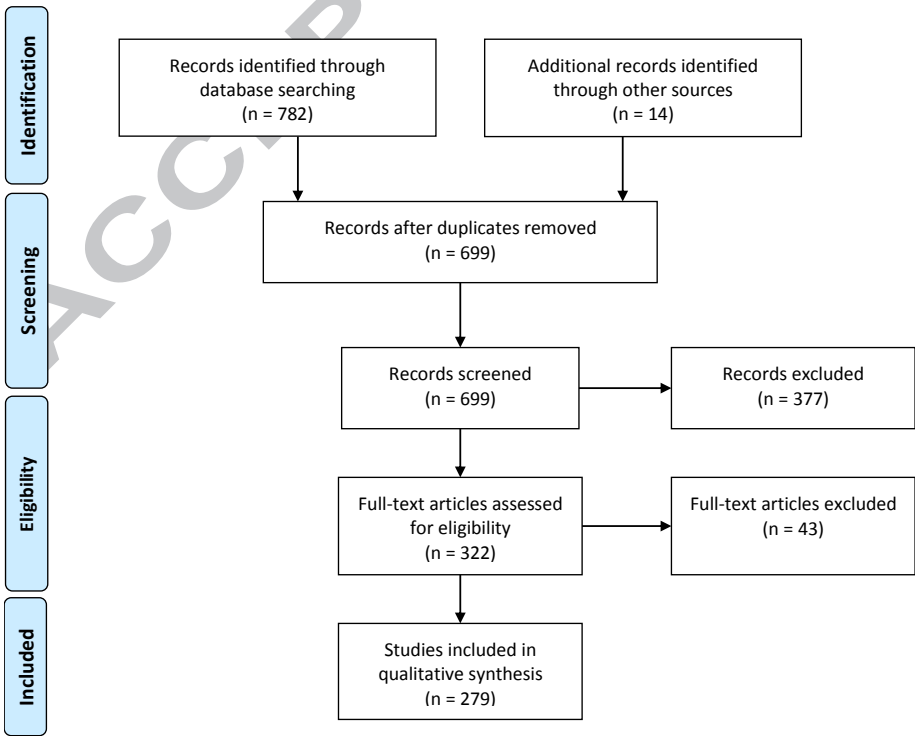

Figure 1 - Article selection flow diagram

In the following of this section we analyze the selected articles along four main dimensions: data types considered in the calculation of the similarity, clinical domains of application, data analysis methods, and translational stage of the reported findings.

3.1 Data types

The analysis of the articles in our review highlighted how different types of data are employed, often in combination, to define similarity between patients. The following five, non-mutually-exclusive, categories were defined: clinical data, molecular data, imaging and biosignals, lab results, and patient-reported outcomes.

The most represented category is the one of molecular data (128 articles, see Fig. 2). This finding is coherent with one of the foundational promises of precision medicine, consisting in the use of molecular data at the point of care to directly impact patient treatment and clinical decision-making. Articles in the molecular data category include studies employing patient similarity measures based on genomics, transcriptomics or metabolomics data, and several approaches where an integration of multiple types of molecular data is performed[4–8].

Interestingly, a relevant number of articles (52) in our review use an integrated approach combining different types of data to define novel patient subgroups that are clinically meaningful. Most of these studies take advantage of clinical data (on which most of the available guideline-defined subgroups already rely) in combination with the deeper molecular profiling[9–11] that precision medicine approaches have been promoting. Recent literature[12] has also pointed out how linking molecular measurement to patient clinical data via electronic medical records should be regarded as the future of precision medicine research. Some studies proposing data integration approaches have also been exploring the possibility of augmenting the traditionally collected clinical data with patient reported outcomes, especially in conditions like chronic back pain or psychiatric diseases where direct symptom reporting by patients is essential to follow the progression of the disease[13–15].

The third most frequent category (50 papers) is represented by more traditional approaches considering clinical data to measure patient similarity. Finally, an interesting trend (19 articles) that has been observed in our review consists in studies directly employing quantitative data from diagnostic imaging (e.g.

MRI[16,17], CT scans[18] and mammograms[19]) or biosignals (e.g. EEGs[20] or ECGs[21]) to group patients and define disease subtypes having different prognosis or different response to treatment.

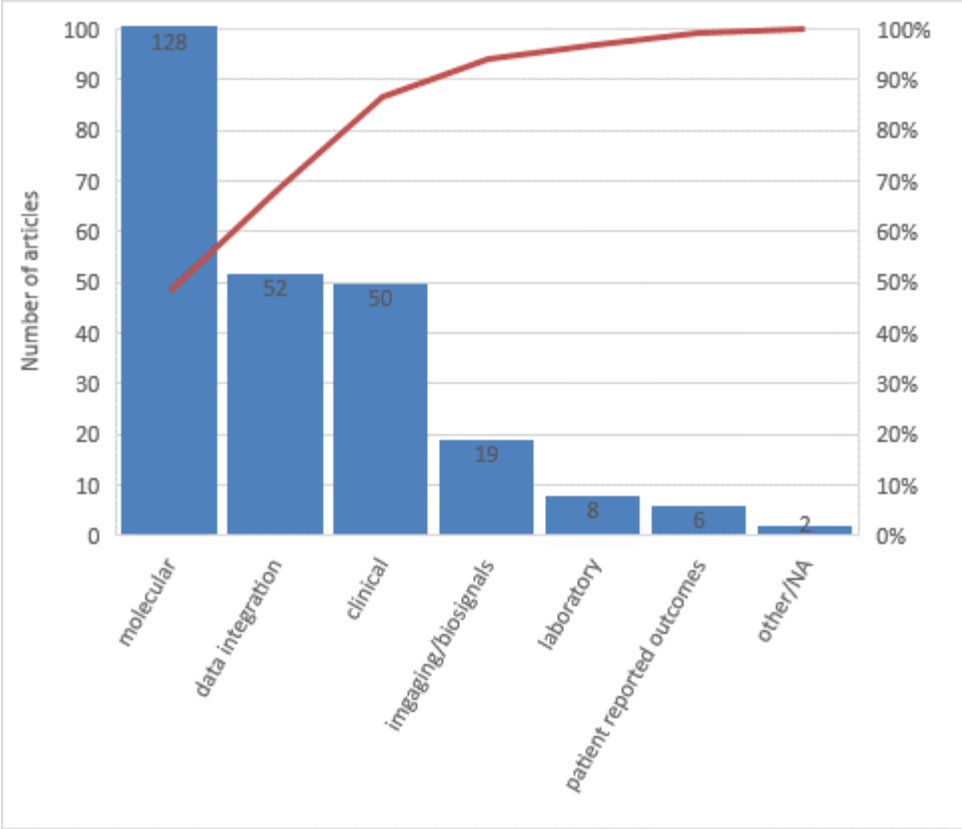

Figure 2 – Number of articles classified according to the types of data used to compute patient similarity. Articles including any combination of data types are counted in the data integration category. Other categories are mutually exclusive. The graph is organized as a Pareto chart, with the orange line showing the cumulative percentage of the total number of articles.

3.2 Clinical domains

When considering domains, we proceeded by labeling each paper according to the major human body apparatus involved in the study, namely *Cardiovascular/Circulatory*, *Digestive/Excretory*, *Endocrine*, *Integumentary/Exocrine*, *Lymphatic/Immune*, *Muscular/Skeletal*, *Nervous/Mental health*, *Renal/Urinary*, *Reproductive*, and *Respiratory* systems. After considering the collected corpus of literature we also added the following labels: *Cancer* (including *pancancer*, i.e. studies considering samples from any kind of cancer), *Sleep*, *Pediatric/juvenile*, *Pain* (if the study is aimed at pain management), *Pathogen/Infection/Microorganism*, *Treatment* (for studies focusing on specific drugs or treatments), *Hospital/ER* (for methods focusing on the management of patient hospitalization) or *Method* (for studies elaborating novel techniques). About 80% of the papers have at least two labels, and about 27% have three or more. Not surprisingly, cancer is the most frequently considered condition, with more than half of the

works (~52%) about either a specific cancer type, such as lung (11 papers), ovarian (10 papers) or colorectal cancer (11 papers), hepatocellular carcinoma (3 papers), or *pancancer* (17 papers).

The most studied cancer type for patient similarity (36 papers) is breast cancer. This drives the category Integumentary/Exocrine system to be the fourth most represented, after nervous system and mental health (25% of the works), and, interestingly, the presentation of a novel approach (25% of the papers). Focusing on mental diseases such as bipolar disorder (2 papers), Parkinson's (7 papers), Alzheimer's (2 papers), Hungtington's (2 papers), schizophrenia (3 papers), or depression (3 papers), might reflect the complexity of classifying mental diseases based on symptoms. Also, tumors affecting the nervous systems (glioblastoma, glioma, brain tumors in general: 19 papers) are well represented in our samples, while amyotrophic lateral/multiple/systemic sclerosis (4 papers) represent a minor focus of studies about patient similarity.

Another major investigation area for patient similarity is lung/respiratory diseases. Several works, for example, consider asthma (11 papers), or chronic obstructive pulmonary disease (COPD, 9 papers).

Interestingly, we found also evidence of an interest in patient similarity for sleep apnoea (2 papers). In summary, besides the predictable major role of cancer (Figure 3), chronic diseases (e.g. asthma, lupus, COPD) play a pivotal role in patient similarity-related investigations.

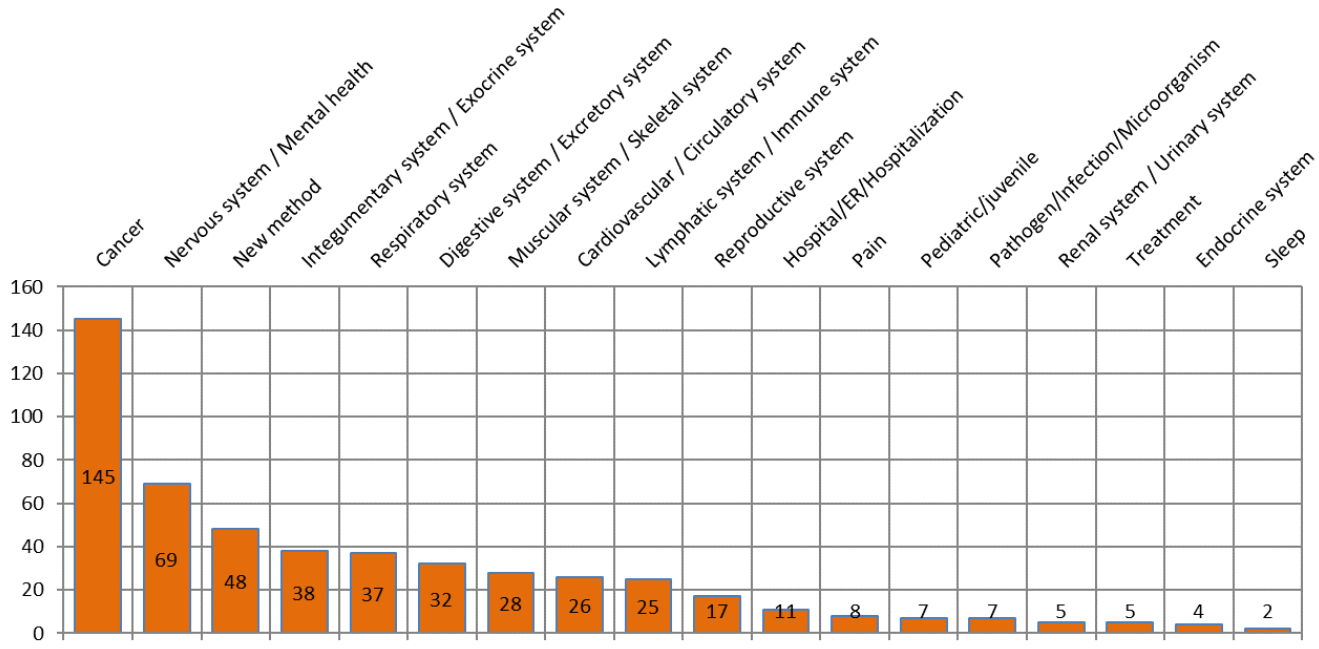

Figure 3 – Number of articles classified according to the clinical domain

3.3 Data analysis methods

One important dimension that was considered in our study regards the methodologies exploited in the papers.

We first divided articles into two main categories: those that did exploit existing methodologies for knowledge discovery in a specific application, and those that instead present a novel approach that is then applied to solve a specific task. Interestingly, out of the 279 considered articles, 66 dealt with the presentation of a novel methodological approach.

As far as traditional methods are concerned, the most popular in our search is, clustering. Clustering techniques are used by 157 papers. Most frequently, clustering is used for processing high dimensional molecular data with the ultimate goal of sub typing diseases. Among the available techniques, the most frequently used are hierarchical and k-means clustering (90 and 22 papers respectively). A few number of applications explore more sophisticated techniques, such as consensus clustering [17,22–28], or model based [8,29–33] approaches.

Besides clustering, other frequently used techniques are those addressing dimensionality reduction, using methodologies such as Principal Component Analysis (PCA), factor analysis, or more complex techniques such as matrix factorization [24,34–36].

Methods falling in the broad categories of clustering and dimensionality reduction represent the vast majority (~95%) of the traditional methods employed in the analyzed papers. Some sporadic examples of different techniques are however present. These include, for example, articles that employ group-based trajectory modeling[37] and already-defined similarity measures such as Jiang's information theoretic similarity[38] or the disease state index[9].

To classify the methodologies proposed in the 66 papers that present novel approaches, we have identified 5 categories, based on the main novelty of the performed research. In particular, the following categories were defined:

- Clustering: when the developed methodology has the main goal of creating groups of patients with similar disease evolution.
- Dimensionality reduction: when the main novelty is the selection of representative features to characterize specific groups of patients. Papers in this group can either present a single feature

selection approach, or a feature selection approach which is then coupled to clustering to obtain groups of similar patients.

- Similarity: when the main goal is to define a novel measure of similarity among patients.
- Software tools: when the main contribution is the availability of a software solution implementing the prosed approach.
- Combination of clustering or similarity metrics and supervised approaches: when the features that characterize group of patients are then used to solve a classification task.

Figure 4 shows the number of papers which use a novel approach that are included in each category in our review.

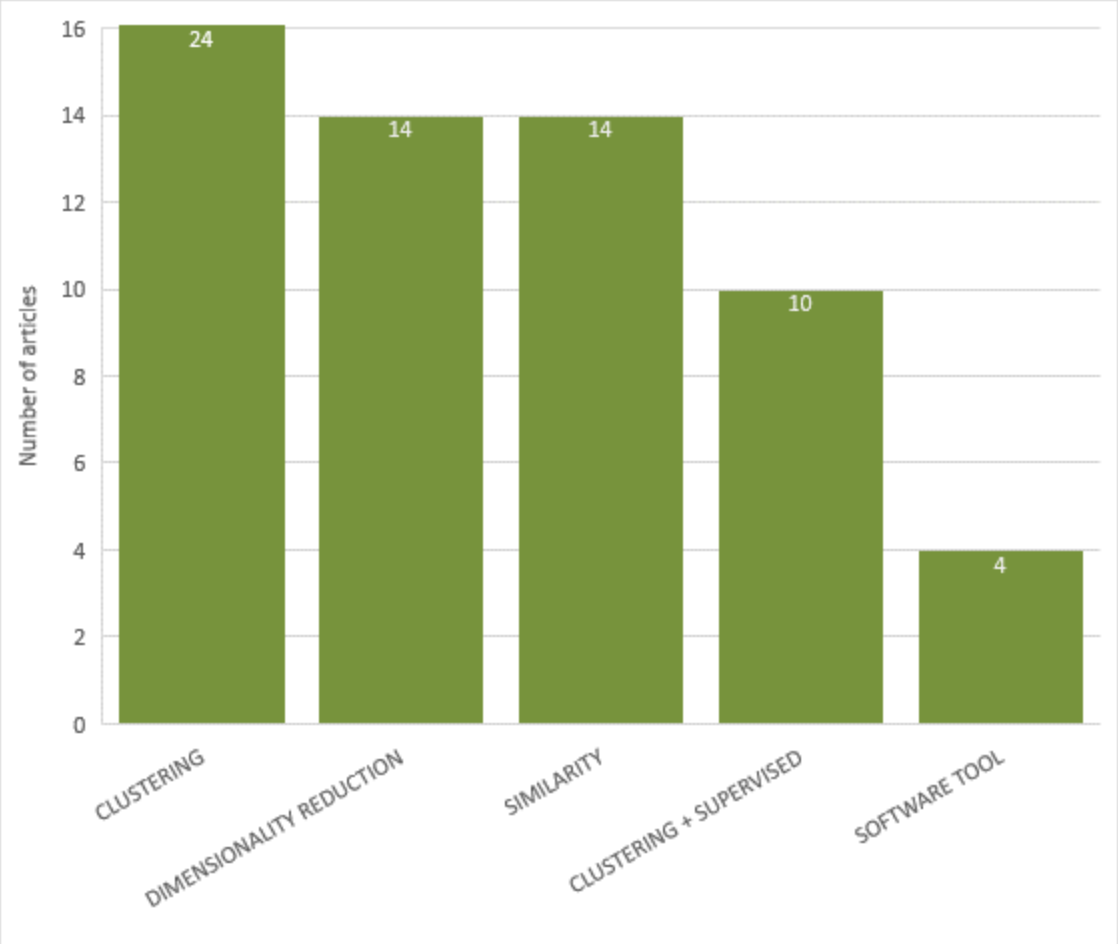

Figure 4 – Number of articles employing a novel approach classified according to the data analysis methods employed for similarity

Novel clustering methods were the most frequent in our selection of papers dealing with novel methodologies. This was expected, as the identification of sub-groups of individuals who can better specify

some characteristic of a disease is often the main application of precision medicine. The 24 papers that were identified by our search exploit a variety of techniques to identify the final clusters. Such techniques include, among others, consensus clustering[39], bi-clustering[40], topological models[41–43], clustering of longitudinal data[44], and voxel clustering[45]. For example, Wang et al[39] define a novel consensus clustering method, called Molecular Regularized Consensus Patient Stratification (MRCPS) to automatically cluster both numerical and categorical data using a number of similarity metrics. The algorithm is based on optimization techniques with regularization, and is applied for validation to mRNA and miRNA expression profiles from primary tumor of breast cancer patients extracted from The Cancer Genome Atlas (TCGA)[46]. Li et al[42] present an interesting framework based on topological analysis to identify type 2 diabetes subgroups. The approach is based on molecular and EMR-extracted clinical data, and it is focused on the construction of a patient-patient similarity network. Highly similar patients are grouped into clusters, which are then compared in terms of clinical features.

The concept of similarity defined in the papers classified in this category is broad and multifaceted. In particular, we found that the proposed algorithms refer to similarity among patients in terms of clinical, diagnostic and/or treatment patterns[11,47–50], similarity among expression profiles computed using molecular data[38,51–53], similarity among time series or signals in general[21,54–56], and even to similarities among textual reports for context-based image retrieval[57].

Similarity techniques and clustering methods have been sometimes coupled to supervised techniques, in different contexts. An interesting example of joint use of supervised and unsupervised approaches can be found in Zhang and Kodell[58], where the authors focus on the identification of patients groups where an ensemble classifier based on a convex-hull, selective-voting algorithm, works well, and of groups where instead the classification performance is poor. This method ultimately assigns to an unknown example a confidence level that expresses the estimated predictive accuracy of the ensemble method applied to the example.

Interestingly, four of the papers classified as presenting a new methodology are software tools. In general, they make a set of methodologies available to users, with a specific attention on the visualization techniques. An example of such tool is Integrated Clustering of Multi-dimensional biomedical data (ICM), presented by He et al[59]. This tool provides a graphical user interface with functionalities related to data fusion, cluster

analysis, and visualization. The available clustering methods are selected to deal with multidimensional data, and are: Similarity Network Fusion (SNF)[60], integrative clustering (iCluster)[61], and concatenation based on spectral clustering.

As for ICM, an interesting aspect that is touched by some the extracted papers is the one related to data integration, which often involves dimensionality reduction techniques. This was specifically addressed in 8 papers out of the 66 proposing novel approaches[4,62–68]. Among these, Planey and Gevaert[64] present CoINcIDE, a framework to discover patients subgroups across multiple datasets. Disease subtypes are obtained by performing a meta-clustering on the nodes of a network, where each of the nodes represents a consensus among the clusters obtained on each specific dataset. In this paper, microarray gene expression datasets related to breast cancer patients were used and integrated. Multi-omic data are considered also by Taskesen et al[63], who identify similarity among patients experiencing 19 types of cancer using a multi-omic bidimensional map. Gligorijević et al[62] present a data fusion approach based on matrix tri-factorization that achieves patients stratification using clinical, molecular, and drug treatment data.

The novel methods proposed in the 66 identified papers present various degree of portability, which vary from general algorithms, to highly specific ones, which make sense only in the context of the disease they are designed to analyze. For example, the aforementioned method proposed by Li and coworkers[42] is a general method based on a new distance formula to be applied to data as heterogeneous as electronic medical records (EMRs) and genotyping single nucleotide variations (SNPs). Intuitively, the very same approach could be ported to other data sets, focused on different diseases, as long as EMR and SNPs data are available. Another general method is reported by SungHwan and coworkers[6]. Here, multiple data sources such as mRNA, miRNA and EMRs are all considered at the same time for patient similarity calculation. Authors apply their method to two different diseases, COPD and interstitial lung disease, to show its portability. On the other hand, an example of disease-specific method is provided by Toddenroth and coworkers[69], with the design of a statistical algorithm exploiting prior medical knowledge to cluster patients suffering from cardiovascular instability. Their method depends on EMR cardiac time series, and it is restricted to analyze solely cardiovascular diseases.

3.4 Translational stage of study findings

A further dimension used to classify the articles in our review consists in the translational stage and practical impact of the reported findings. We identified three levels of maturity: treatment targeting (refinement of a population target for a specific treatment), outcome-based subgroup discovery (investigating reasons behind different observed outcomes or survival in patients with similar clinical presentation), identification of disease subgroup. In the following we discuss each of the three, starting from treatment targeting.

One of the highest priorities of precision medicine research is to improve treatment targeting strategies for specific pathologies, which may ultimately improve medication effectiveness and clinical outcomes. However only a limited number of studies (29 articles) that involved patient similarity and precision medicine were mature enough to produce knowledge that directly informs treatment targeting decisions. The majority of studies that are able to achieve this belong to the cancer domain. Molecular profiling of cancer has been consistently performed in latest years and advances in cancer care are the most paradigmatic examples of the benefits of precision medicine research. Breast cancer is one of the instances where knowledge of specific sub-phenotypes (e.g. triple negative BC) is already driving treatment choices in clinical practice[70], for example on the benefits of neo-adjuvant chemotherapy[23]. Similarly, research on other types of cancer, where molecular characterization of disease phenotypes is already at an advanced stage, often focus on the effectiveness of a specific drug on a well-defined sub-population of patients. This is the case, for example, of ovarian cancer[71] or hepatocellular carcinoma[72] where specific studies have been conducted aiming at identifying actionable drug targets or assessing effectiveness of DNA-repair adjuvant drugs. It is not uncommon for this type of studies to even focus on a single chemical, like bortezomib and its application in myeloma chemotherapy[73], and provide recommendations on how to maximize its effectiveness with a careful selection of patients. Even if cancer care has been one of the leading applications since the inception of precision medicine, also other disease types are well represented despite their molecular profiling may not be at an equally advanced stage yet. Indeed, drug targeting strategies that take advantage of data other than molecular have also proven effective in improving outcomes for patients. For example neurological disorders like depression[74] and social anxiety disorder[75], where the selection of an effective pharmacological therapy is often non-trivial, may benefit from the definition of patient sub-groups. In the study from Stein and colleagues[75], this definition is based on a similarity

measure defined using clinical data and patient reported outcomes, integrated with genetic characterization where available. Finally, our review identified another noteworthy trend consisting of articles employing a similarity score defined on diagnostic images to implement an automatic treatment plan selection, usually for radiotherapy. Some of the most represented domains in this line of research are the ones where radiotherapy plays a pivotal role in cancer treatment like head and neck cancer[76], and bladder cancer[77].

Not all types of cancer have an equally advanced understanding of the molecular mechanisms driving pathogenesis and disease progression as BC. In these cases, different outcomes are nonetheless regularly observed in patients with rather similar clinical presentation: studies deepening this aspect are included in our outcome-based subgroup discovery category (53 articles). Researchers often investigate whether different phenotypes of the same disease do exist, but are still undescribed. Many of these studies, as in the case of squamous cell carcinoma[8], use the integration of multiple –omics in order to explore associations between novel disease sub-groups and the observed variability in terms of clinical outcomes. Some interesting research efforts[78] also focus on advancing well-established cancer staging classification schemes, like TNM, to increase sub-stratification of patients and identify and separate favorable and unfavorable outcomes. Moreover, our analyses have identified methodological and domain-independent contributions that exploit patient similarity to predict disease prognosis[49], thus highlighting that these approaches have the potential to generalize to a wide range of medical domains and diseases.

In clinical application domains where specific targeted therapies are not yet available, and where differences in clinical outcomes between different patients are not emergent, patient similarity still finds its role in discovering new basic knowledge related to biological mechanisms underlying the diseases and their etiology. These studies are included in the identification of disease subgroup category and are the most numerous in our review (174 articles). Diseases like depression, bipolar disorder and schizophrenia, where the diagnosis is particularly challenging, all end up in this category. For these conditions, subgroup identification is improved by defining biosignatures to avoiding confusion between competing diagnoses[79]. Other studies, especially on complex chronic conditions such as asthma[80] and COPD[81] or obstructive sleep apnea[82], aim at defining novel sub-phenotypes for better characterization of these highly heterogeneous diseases. Despite the fact that these studies may not directly deal with therapy targeting or are not immediately informative for prognosis, they remain among the most important needs that this early-stage

research aims at fulfilling. Indeed, also cancer research still has so many significant challenges that a proper combination of biological, molecular and computational methodologies is essential[35]. The need to enable and facilitate such kind of exploratory analyses is also evident from current research in visual analytics[83] and integrated clustering[59] of patient data for precision medicine.

4 Discussion

In this article we presented a systematic review of the literature about the use of patient similarity measures in precision medicine research. A previous work from Brown[84] has discussed a similar subject, but, to our best knowledge, the present article is the first systematic literature review on the topic.

Our findings highlight how patient similarity is, and is likely to remain, a central topic in precision medicine research. Applications to the cancer domain are the most represented in our review and the majority of the studies we analyzed rely on well-established data mining approaches like clustering and dimensionality reduction. A portion of the studies (66, see methods sub-section within Results) reported novel methodological contributions, only a minority of which have the potential to generalize to other medical domains of application. These findings point out how a large part of the research currently being carried out in precision medicine and employing patient similarity is still focused on answering narrowly defined research questions that might improve outcomes for specific patients which currently have poor prognosis (e.g. drug repurposing studies, refined treatment targeting based on molecular profiling, etc). The most important findings reported in these kind of studies will eventually be incorporated in future updates of clinical practice guidelines. However, barriers including infrequent updates of guidelines, slow dissemination and adoption of such updates in clinical practice and the limited number of patients that are directly affected by a specific new recommendation is likely to affect the net translational impact of such research[85].

On the other hand, patient similarity and precision medicine have the potential of making a difference for a much broader population of patients, which is not limited to the ones enrolled in the most advanced cancer centers where molecular profiling and advanced therapies are available. This is true especially considering that our study highlighted how chronic diseases are indeed a well-represented domain.

The WHO ranked the leading causes of Disability-Adjusted Life Years (DALYs, the unit utilized by WHO to measure the impact of mortality and loss of health due to diseases and injuries; roughly, one DALY can be thought of as one lost year of healthy life), as well as the causes of death [86]. We analyzed the first 10

causes of both DALYs and death in order to compare the current state of patient-similarity oriented precision medicine with the world- causes. These results are summarized in Table 1.

Table 1

Cause	Rank in causes of DALY	Rank in causes of death	Accounts for % of DALYs	Accounts for % of deaths	works directly addressing the cause in survey
Ischaemic heart disease	1st	1st	~12%	~25%	4/279 (1.43%)
Stroke	3rd	2nd			
Lower respiratory infections	2nd	3rd	~9%	~11.3%	8/279 (2.87%)
Chronic obstructive pulmonary disease	7th	4th			
Preterm birth complications	4th	14h	3.8%	1.9%	1/279 (0.35%)
Diabetes Mellitus	8th	6th	2.6%	2.8%	0.07% (2/729)
Trachea, bronchus, lung cancers	17th	5th	1.5%	3%	11/279 (3.94%)
Alzheimer disease and other dementias	-	7th	<1.2%	2.7%	2/729 (0.07%)
Diarrhoeal diseases	5th	8th	3.2%	2.5%	none
Birth asphyxia and birth trauma	9th	20th	2.5%	1.2%	
Tubercolosis	12th	-	2.1%	<1.2%	
Congenital anomalies	10th	-	2.4%	<1.2%	

Excluding the one non-medical causes (road injury, ranked 6th and 10th, respectively, for DALYs and death), we are left with 9 causes. The first cause for both DALYs and death is "Ischaemic heart disease", which, along with "Stroke" (3rd cause of DALY and 2nd cause of death) accounts for about 12% of the total DALYs, and about 25% of the deaths. Our review shows that studies on Cardiovascular / Circulatory system accounts for about the 10% of the total, with only 2 works directly accounting for heart failure, and other 2 for myocardial infarction. "Lower respiratory infections" and "Chronic obstructive pulmonary disease" (COPD) are, respectively, the 2nd and 7th causes of DALYs, and the 3rd and 4th causes of death. Although we did not find any study addressing lower respiratory infections (we have only 2 studies about sepsis), there are 8 papers on COPD. Preterm birth complications represent the 4th DALY cause. We found 1 work addressing it explicitly. Diabetes mellitus represents the 8th cause of DALY and 6th cause of death. We found 2 papers addressing it explicitly. "Alzheimer disease and other dementias" are reported as the 7th cause of death. We

found 2 papers explicitly addressing Alzheimer's disease. For "Diarrhoeal diseases" (5th cause of DALYs and 8th cause of death), "Birth asphyxia and birth trauma" (9th cause of DALYs), "Tubercolosis" (9th cause of death), and "Congenital anomalies" (9th cause of DALYs) we did not find any specific paper. Cancer is present as 5th cause of death with "Trachea, bronchus, lung cancers". We found 11 papers in our survey focusing on lung cancer. We found patient-similarity precision medicine research partially overlapping with the leading causes of DALYs and death. Unsurprisingly, research seems to be more unbalanced towards diseases that are massively present in both low- and high-development countries (e.g., lung cancer[87] or diabetes[88]) , while typical diseases related to low-development countries, such as diarrhoeal diseases[89], do not appear in the retrieved literature.

We believe that an interesting direction for future research efforts would be represented by the integration of patient similarity measures into decision support systems, as it might accelerate the uptake of precision medicine research findings[90,91]. To this end, an encouraging fact is that our review identified a small number (4) of software tools offering similarity-enabled capabilities like patient cluster analysis, integration of heterogeneous data, and visualization that constitute a common, domain-independent toolset for precision medicine research. Furthermore, the presence of a small but significant set of studies which developed portable novel methodological contributions regarding patient similarity is also likely to contribute to this trend. A recent review article by Sharafoddini[92] highlighted how two chronic conditions, namely cardiovascular disease and diabetes, were the main focus of computer-based approaches for predicting patients' future health status based on health data and patient similarity. The same article highlighted that patient similarity-based modeling outperformed population-based predictive methods in two studies.

Our work has some limitations. Our search was limited to three databases, namely PubMed, Scopus and IEEE. In addition, we limited our temporal scope to research published in the last 6 years. This, together with the eligibility criteria we defined, might have resulted in significant research being left out from our review. Interestingly, our searches did not retrieve any paper exploiting the concept of kinship matrices[93]. Kinship matrices are well known in genomics[94,95], and represent an effective approach to dimensionality reduction. Intuitively, given *n* samples with *p* measurements (e.g. patients and their SNPs), kinship matrices are *n x n* matrices accounting for the fraction of equal alleles for each possible patient pair. Therefore, they embed the concept of patient similarity, genetically speaking, by construction. Despite the availability of

related literature[94,96,97], it seems the integration of kinship matrices in patient subgroup identification methods is still underrepresented in the literature.

Regarding data analysis methodologies, we have found only a few contributions related to deep learning. Deep learning is currently a hot topic in machine learning and artificial intelligence research. The outstanding results obtained by deep neural networks in fields like image analysis and their ability of handling massive volumes of data through embedded dimensionality reduction have also attracted the attention of researchers in the medical field. Several works, in fact, exploit data fusion via deep learning to face the variety and complexity of medical data. For example Miotto and colleagues[98] recently reported on a method for devising compact representations of patients using a three-layer stack of denoising autoencoders, which outperformed those based on raw EHR data on a health state prediction task. Similarly, an approach based on deep belief networks has been applied by Liang and colleagues[99] in 2014 to identify meaningful disease subtypes from multi-platform molecular cancer data. Finally, an example of how deep learning with convolutional neural networks applied to chest CT scans can predict 5-year survival has been recently reported by Oakden-Rayner and colleagues[18].

As a final remark, we did not retrieve any instance of case-based methods applied to precision medicine. Given the close relationship, both methodological and historical[100], between case-based approaches and patient similarity, we expected to find relevant literature in our search. For example, recent articles describe the application of case-based reasoning to the definition of patient subgroups in breast cancer[101] or lung cancer[102]. This highlights a possible limitation of our search strategy, which was not able to retrieve such articles, probably due to the fact that they implicitly include the concept of patient similarity in their core methodology without explicitly mentioning it as a keyword or in title and abstract of the article.

5 Conclusion

This article presented a systematic review of the literature regarding the use of patient similarity to enable precision medicine research. Two hundred seventy-nine articles were analyzed and classified according to four dimensions: data types considered in the calculation of the similarity, clinical domains of application, data analysis methods, and translational stage of the reported findings. Cancer-related research employing molecular profiling and standard data analysis techniques such as clustering constitutes a relevant portion of

the retrieved studies. However, also chronic and psychiatric diseases show a relevant number of works published in the field. Interestingly, almost one quarter of the analyzed studies presented a novel methodology involving patient similarity for the identification of clinically-meaningful subgroups. Among these, the most advanced employ data integration strategies and are able to generalize to different clinical domains. Integration of such novel techniques in software tools devoted to medical decision support constitutes and interesting trend for future precision medicine research, which has the potential to accelerate its translational impact.

References

[1] E. Steinberg, S. Greenfield, D.M. Wolman, M. Mancher, R. Graham, Clinical practice guidelines we can trust, National Academies Press, 2011.

[2] F.S. Collins, H. Varmus, A New Initiative on Precision Medicine, N. Engl. J. Med. 372 (2015) 793–795. doi:10.1056/NEJMp1500523.

[3] J.L. Jameson, D.L. Longo, Precision medicine--personalized, problematic, and promising, N. Engl. J. Med. 372 (2015) 2229–2234. doi:10.1056/NEJMsb1503104.

[4] A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, D. Greco, MVDA: a multi-view genomic data integration methodology, BMC Bioinformatics. 16 (2015) 261. doi:10.1186/s12859-015-0680-3.

[5] H. Ross-Adams, A.D. Lamb, M.J. Dunning, S. Halim, J. Lindberg, C.M. Massie, L.A. Egevad, R. Russell, A. Ramos-Montoya, S.L. Vowler, N.L. Sharma, J. Kay, H. Whitaker, J. Clark, R. Hurst, V.J. Gnanapragasam, N.C. Shah, A.Y. Warren, C.S. Cooper, A.G. Lynch, R. Stark, I.G. Mills, H. Grönberg, D.E. Neal, CamCaP Study Group, Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study, EBioMedicine. 2 (2015) 1133–1144. doi:10.1016/j.ebiom.2015.07.017.

[6] S. Kim, J.D. Herazo-Maya, D.D. Kang, B.M. Juan-Guardela, J. Tedrow, F.J. Martinez, F.C. Sciurba, G.C. Tseng, N. Kaminski, Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes, BMC Genomics. 16 (2015) 924. doi:10.1186/s12864-015-2170-4.

[7] J.B. Andersen, B. Spee, B.R. Blechacz, I. Avital, M. Komuta, A. Barbour, E.A. Conner, M.C. Gillen, T. Roskams, L.R. Roberts, V.M. Factor, S.S. Thorgeirsson, Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors, Gastroenterology. 142 (2012) 1021-1031.e15. doi:10.1053/j.gastro.2011.12.005.

[8] A.C. Jung, S. Job, S. Ledrappier, C. Macabre, J. Abecassis, A. de Reyniès, B. Wasylyk, A poor prognosis subtype of HNSCC is consistently observed across methylome, transcriptome, and miRNome analysis, Clin. Cancer Res. 19 (2013) 4174–4184. doi:10.1158/1078-0432.CCR-12-3690.

[9] A. Hall, J. Mattila, J. Koikkalainen, J. Lötjonen, R. Wolz, P. Scheltens, G. Frisoni, M. Tsolaki, F. Nobili, Y. Freund-Levi, L. Minthon, L. Frölich, H. Hampel, P.J. Visser, H. Soininen, Predicting progression from cognitive impairment to alzheimer's disease with the disease state index, Current Alzheimer Research. 12 (2015) 69–79.

[10] A. Sadanandam, C.A. Lyssiotis, K. Homicsko, E.A. Collisson, W.J. Gibb, S. Wullschleger, L.C.G. Ostos, W.A. Lannon, C. Grotzinger, M. Del Rio, B. Lhermitte, A.B. Olshen, B. Wiedenmann, L.C. Cantley, J.W. Gray, D. Hanahan, A colorectal cancer classification system

that associates cellular phenotype and responses to therapy, Nature Medicine. 19 (2013) 619–625. doi:10.1038/nm.3175.
[11] G.S. Ow, Z. Tang, V.A. Kuznetsov, Big data and computational biology strategy for personalized prognosis, Oncotarget. 7 (2016) 40200–40220. doi:10.18632/oncotarget.9571.
[12] A.J. Vargas, C.C. Harris, Biomarker development in the precision medicine era: lung cancer as a case study, Nat Rev Cancer. 16 (2016) 525–537. doi:10.1038/nrc.2016.56.
[13] S.E. Reme, W.S. Shaw, I.A. Steenstra, M.J. Woiszwillo, G. Pransky, S.J. Linton, Distressed, immobilized, or lacking employer support? A sub-classification of acute work-related low back pain, J Occup Rehabil. 22 (2012) 541–552. doi:10.1007/s10926-012-9370-4.
[14] O. Hirsch, K. Strauch, H. Held, M. Redaelli, J.-F. Chenot, C. Leonhardt, S. Keller, E. Baum, M. Pfingsten, J. Hildebrandt, H.-D. Basler, M.M. Kochen, N. Donner-Banzhoff, A. Becker, Low back pain patient subgroups in primary care: pain characteristics, psychosocial determinants, and health care utilization, Clin J Pain. 30 (2014) 1023–1032. doi:10.1097/AJP.0000000000000080.
[15] T. Tarpey, E. Petkova, L. Zhu, Stratified Psychiatry via Convexity-Based Clustering with Applications Towards Moderator Analysis, Stat Interface. 9 (2016) 255–266. doi:10.4310/SII.2016.v9.n3.a1.
[16] K.H. Brodersen, L. Deserno, F. Schlagenhauf, Z. Lin, W.D. Penny, J.M. Buhmann, K.E. Stephan, Dissecting psychiatric spectrum disorders by generative embedding, Neuroimage Clin. 4 (2014) 98–111. doi:10.1016/j.nicl.2013.11.002.
[17] T.T. Liu, A.S. Achrol, L.A. Mitchell, S.A. Rodriguez, A. Feroze, Michael Iv, C. Kim, N. Chaudhary, O. Gevaert, J.M. Stuart, G.R. Harsh, S.D. Chang, D.L. Rubin, Magnetic resonance perfusion image features uncover an angiogenic subgroup of glioblastoma patients with poor survival and better response to antiangiogenic treatment, Neuro-Oncology. (2016). doi:10.1093/neuonc/now270.
[18] L. Oakden-Rayner, G. Carneiro, T. Bessen, J.C. Nascimento, A.P. Bradley, L.J. Palmer, Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework, Scientific Reports. 7 (2017) 1648. doi:10.1038/s41598-017-01931-w.
[19] P. Casti, A. Mencattini, M. Salmeri, R.M. Rangayyan, Analysis of Structural Similarity in Mammograms for Detection of Bilateral Asymmetry, IEEE Transactions on Medical Imaging. 34 (2015) 662–671. doi:10.1109/TMI.2014.2365436.
[20] M. Kaur, J. Lagopoulos, P.B. Ward, T.L. Watson, S.L. Naismith, I.B. Hickie, D.F. Hermens, Mismatch negativity/P3a complex in young people with psychiatric disorders: a cluster analysis, PLoS ONE. 7 (2012) e51871. doi:10.1371/journal.pone.0051871.
[21] J. Park, K. Kang, HeartSearcher: finds patients with similar arrhythmias based on heartbeat classification, IET Systems Biology. 9 (2015) 303–308. doi:10.1049/iet-syb.2015.0011.
[22] L. Marisa, A. de Reyniès, A. Duval, J. Selves, M.P. Gaub, L. Vescovo, M.-C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J.-F. Fléjou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig, V. Boige, Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value, PLoS Med. 10 (2013) e1001453. doi:10.1371/journal.pmed.1001453.
[23] H. Masuda, K.A. Baggerly, Y. Wang, Y. Zhang, A.M. Gonzalez-Angulo, F. Meric-Bernstam, V. Valero, B.D. Lehmann, J.A. Pietenpol, G.N. Hortobagyi, W.F. Symmans, N.T. Ueno, Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes, Clin. Cancer Res. 19 (2013) 5533–5540. doi:10.1158/1078-0432.CCR-13-0799.
[24] H. Chen, J. Xu, J. Hong, R. Tang, X. Zhang, J.-Y. Fang, Long noncoding RNA profiles identify five distinct molecular subtypes of colorectal cancer with clinical relevance, Molecular Oncology. 8 (2014) 1393–1403. doi:10.1016/j.molonc.2014.05.010.

[25] M. Planck, S. Isaksson, S. Veerla, J. Staaf, Identification of transcriptional subgroups in EGFR-mutated and EGFR/KRAS wild-type lung adenocarcinoma reveals Gene signatures associated with patient outcome, Clinical Cancer Research. 19 (2013) 5116–5126. doi:10.1158/1078-0432.CCR-13-0928.

[26] M.R. Aure, V. Vitelli, S. Jernström, S. Kumar, M. Krohn, E.U. Due, T.H. Haukaas, S.-K. Leivonen, H.K.M. Vollan, T. Lüders, E. Rødland, C.J. Vaske, W. Zhao, E.K. Møller, S. Nord, G.F. Giskeødegård, T.F. Bathen, C. Caldas, T. Tramm, J. Alsner, J. Overgaard, J. Geisler, I.R.K. Bukholm, B. Naume, E. Schlichting, T. Sauer, G.B. Mills, R. Kåresen, G.M. Mælandsmo, O.C. Lingjærde, A. Frigessi, V.N. Kristensen, A.-L. Børresen-Dale, K.K. Sahlberg, E. Borgen, O. Engebråten, O. Fodstad, B. Fritzman, O. Garred, G.A. Geitvik, A. Langerød, S. Hofvind, H.G. Russnes, H.K. Skjerven, T. Sørlie, OSBREAC, Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome, Breast Cancer Research. 19 (2017). doi:10.1186/s13058-017-0812-y.

[27] B.P. Scicluna, L.A. van Vught, A.H. Zwinderman, M.A. Wiewel, E.E. Davenport, K.L. Burnham, P. Nurnberg, M.J. Schultz, J. Horn, O.L. Cremer, M.J. Bonten, C.J. Hinds, H.R. Wong, J.C. Knight, T. van der Poll, Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study., Lancet Respir Med. 5 (2017) 816–826. doi:10.1016/S2213-2600(17)30294-1.

[28] K.C. Vranas, J.K. Jopling, T.E. Sweeney, M.C. Ramsey, A.S. Milstein, C.G. Slatore, G.J. Escobar, V.X. Liu, Identifying Distinct Subgroups of ICU Patients: A Machine Learning Approach., Crit Care Med. 45 (2017) 1607–1615. doi:10.1097/CCM.0000000000002548.

[29] J.D. Banfield, A.E. Raftery, Model-Based Gaussian and Non-Gaussian Clustering, Biometrics. 49 (1993) 803–821. doi:10.2307/2532201.

[30] R. Chen, J. Sun, R.S. Dittus, D. Fabbri, J. Kirby, C.L. Laffer, C.D. McNaughton, B. Malin, Patient Stratification Using Electronic Health Records from a Chronic Disease Management Program, IEEE J Biomed Health Inform. (2016). doi:10.1109/JBHI.2016.2514264.

[31] M. Bauer, T. Glenn, M. Alda, O.A. Andreassen, E. Angelopoulos, R. Ardau, C. Baethge, R. Bauer, F. Bellivier, R.H. Belmaker, M. Berk, T.D. Bjella, L. Bossini, Y. Bersudsky, E.Y.W. Cheung, J. Conell, M. Del Zompo, S. Dodd, B. Etain, A. Fagiolini, M.A. Frye, K.N. Fountoulakis, J. Garneau-Fournier, A. Gonzalez-Pinto, H. Harima, S. Hassel, C. Henry, A. Iacovides, E.T. Isometsä, F. Kapczinski, S. Kliwicki, B. König, R. Krogh, M. Kunz, B. Lafer, E.R. Larsen, U. Lewitzka, C. Lopez-Jaramillo, G. MacQueen, M. Manchia, W. Marsh, M. Martinez-Cengotitabengoa, I. Melle, S. Monteith, G. Morken, R. Munoz, F.G. Nery, C. O'Donovan, Y. Osher, A. Pfennig, D. Quiroz, R. Ramesar, N. Rasgon, A. Reif, P. Ritter, J.K. Rybakowski, K. Sagduyu, A.M. Scippa, E. Severus, C. Simhandl, D.J. Stein, S. Strejilevich, A. Hatim Sulaiman, K. Suominen, H. Tagata, Y. Tatebayashi, C. Torrent, E. Vieta, B. Viswanath, M.J. Wanchoo, M. Zetin, P.C. Whybrow, Influence of birth cohort on age of onset cluster analysis in bipolar I disorder, European Psychiatry. 30 (2015) 99–105. doi:10.1016/j.eurpsy.2014.10.005.

[32] W. Zhang, H. Feng, H. Wu, X. Zheng, Accounting for tumor purity improves cancer subtype classification from DNA methylation data., Bioinformatics. 33 (2017) 2651–2657. doi:10.1093/bioinformatics/btx303.

[33] Y. Dai, S. Lokhandwala, W. Long, R. Mark, L.-W.H. Lehman, Phenotyping Hypotensive Patients in Critical Care Using Hospital Discharge Summaries., IEEE EMBS Int Conf Biomed Health Inform. 2017 (2017) 401–404. doi:10.1109/BHI.2017.7897290.

[34] S.J. Bradley, A. Suarez-Fueyo, D.R. Moss, V.C. Kyttaris, G.C. Tsokos, T Cell Transcriptomes Describe Patient Subtypes in Systemic Lupus Erythematosus, PLoS ONE. 10 (2015) e0141171. doi:10.1371/journal.pone.0141171.

[35] S. Park, S.-J. Kim, D. Yu, S. Peña-Llopis, J. Gao, J.S. Park, B. Chen, J. Norris, X. Wang, M. Chen, M. Kim, J. Yong, Z. Wardak, K. Choe, M. Story, T. Starr, J.-H. Cheong, T.H. Hwang, An integrative somatic mutation analysis to identify pathways linked with survival outcomes

across 19 cancer types, Bioinformatics. 32 (2016) 1643–1651. doi:10.1093/bioinformatics/btv692.
[36] S. Yepes, M.M. Torres, R.E. Andrade, Clustering of expression data in chronic lymphocytic leukemia reveals new molecular subdivisions, PLoS ONE. 10 (2015). doi:10.1371/journal.pone.0137132.
[37] X.S. Wang, Q. Shi, P.M. Dougherty, C. Eng, T.R. Mendoza, L.A. Williams, D.R. Fogelman, C.S. Cleeland, Prechemotherapy Touch Sensation Deficits Predict Oxaliplatin-Induced Neuropathy in Patients with Colorectal Cancer, Oncology (Switzerland). 90 (2016) 127–135. doi:10.1159/000443377.
[38] V. Gardeux, A.D. Arslan, I. Achour, T.-T. Ho, W.T. Beck, Y.A. Lussier, Concordance of deregulated mechanisms unveiled in underpowered experiments: PTBP1 knockdown case study, BMC Med Genomics. 7 Suppl 1 (2014) S1. doi:10.1186/1755-8794-7-S1-S1.
[39] C. Wang, R. Machiraju, K. Huang, Breast cancer patient stratification using a molecular regularized consensus clustering method, Methods. 67 (2014) 304–312. doi:10.1016/j.ymeth.2014.03.005.
[40] S. Khakabimamaghani, M. Ester, Bayesian biclustering for patient stratification, Pac Symp Biocomput. 21 (2016) 345–356.
[41] M. Pyatnitskiy, I. Mazo, M. Shkrob, E. Schwartz, E. Kotelnikova, Clustering gene expression regulators: new approach to disease subtyping, PLoS ONE. 9 (2014) e84955. doi:10.1371/journal.pone.0084955.
[42] L. Li, W.-Y. Cheng, B.S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E.P. Bottinger, J.T. Dudley, Identification of type 2 diabetes subgroups through topological analysis of patient similarity, Sci Transl Med. 7 (2015) 311ra174. doi:10.1126/scitranslmed.aaa9364.
[43] T. Xu, T.D. Le, L. Liu, R. Wang, B. Sun, J. Li, Identifying cancer subtypes from miRNA-TFmRNA regulatory networks and expression data, PLoS ONE. 11 (2016). doi:10.1371/journal.pone.0152792.
[44] C. Schramm, C. Vial, A.-C. Bachoud-Lévi, S. Katsahian, Clustering of longitudinal data by using an extended baseline: A new method for treatment efficacy clustering in longitudinal data, Stat Methods Med Res. (2015). doi:10.1177/0962280215621591.
[45] E. Schreibmann, A.F. Waller, I. Crocker, W. Curran, T. Fox, Voxel clustering for quantifying PET-based treatment response assessment, Medical Physics. 40 (2013). doi:10.1118/1.4764900.
[46] Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours, Nature. 490 (2012) 61–70. doi:10.1038/nature11412.
[47] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Towards personalized medicine: leveraging patient similarity and drug similarity analytics, AMIA Jt Summits Transl Sci Proc. 2014 (2014) 132–136.
[48] F. Wang, Adaptive semi-supervised recursive tree partitioning: The ART towards large scale patient indexing in personalized healthcare, J Biomed Inform. 55 (2015) 41–54. doi:10.1016/j.jbi.2015.01.009.
[49] F. Wang, J. Sun, PSF: A Unified Patient Similarity Evaluation Framework Through Metric Learning With Weak Supervision, IEEE Journal of Biomedical and Health Informatics. 19 (2015) 1053–1060. doi:10.1109/JBHI.2015.2425365.
[50] Z. Huang, W. Dong, H. Duan, H. Li, Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem, Method, and Applications, IEEE Journal of Biomedical and Health Informatics. 18 (2014) 4–14. doi:10.1109/JBHI.2013.2274281.
[51] V. Gardeux, A. Bosco, J. Li, M.J. Halonen, D. Jackson, F.D. Martinez, Y.A. Lussier, Towards a PBMC “virogram assay” for precision medicine: Concordance between ex vivo and in vivo viral infection transcriptomes, J Biomed Inform. 55 (2015) 94–103. doi:10.1016/j.jbi.2015.03.003.

[52] A.G. Schissler, V. Gardeux, Q. Li, I. Achour, H. Li, W.W. Piegorsch, Y.A. Lussier, Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival, Bioinformatics. 31 (2015) i293-302. doi:10.1093/bioinformatics/btv253.
[53] Q. Li, A.G. Schissler, V. Gardeux, J. Berghout, I. Achour, C. Kenost, H. Li, H.H. Zhang, Y.A. Lussier, kMEn: Analyzing noisy and bidirectional transcriptional pathway responses in single subjects, J Biomed Inform. 66 (2017) 32–41. doi:10.1016/j.jbi.2016.12.009.
[54] S. Segarra, W. Huang, A. Ribeiro, Diffusion and Superposition Distances for Signals Supported on Networks, IEEE Transactions on Signal and Information Processing over Networks. 1 (2015) 20–32. doi:10.1109/TSIPN.2015.2429471.
[55] M. Karg, W. Seiberl, F. Kreuzpointner, J.P. Haas, D. Kulić, Clinical Gait Analysis: Comparing Explicit State Duration HMMs Using a Reference-Based Index, IEEE Transactions on Neural Systems and Rehabilitation Engineering. 23 (2015) 319–331. doi:10.1109/TNSRE.2014.2362862.
[56] J. Dauwels, T. Weber, F. Vialatte, T. Musha, A. Cichocki, Quantifying Statistical Interdependence, Part III: N gt; 2 Point Processes, Neural Computation. 24 (2012) 408–454. doi:10.1162/NECO_a_00235.
[57] J. Ramos, T.T.J.P. Kockelkorn, I. Ramos, R. Ramos, J. Grutters, M.A. Viergever, B. van Ginneken, A. Campilho, Content-Based Image Retrieval by Metric Learning From Radiology Reports: Application to Interstitial Lung Diseases, IEEE Journal of Biomedical and Health Informatics. 20 (2016) 281–292. doi:10.1109/JBHI.2014.2375491.
[58] C. Zhang, R.L. Kodell, Subpopulation-specific confidence designation for more informative biomedical classification, Artif Intell Med. 58 (2013) 155–163. doi:10.1016/j.artmed.2013.04.008.
[59] S. He, H. He, W. Xu, X. Huang, S. Jiang, F. Li, F. He, X. Bo, ICM: a web server for integrated clustering of multi-dimensional biomedical data, Nucleic Acids Res. 44 (2016) W154-159. doi:10.1093/nar/gkw378.
[60] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods. 11 (2014) 333–337. doi:10.1038/nmeth.2810.
[61] R. Shen, A.B. Olshen, M. Ladanyi, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, Bioinformatics. 25 (2009) 2906–2912. doi:10.1093/bioinformatics/btp543.
[62] V. Gligorijević, N. Malod-Dognin, N. Pržulj, Patient-specific data fusion for cancer stratification and personalised treatment, Pac Symp Biocomput. 21 (2016) 321–332.
[63] E. Taskesen, S.M.H. Huisman, A. Mahfouz, J.H. Krijthe, J. de Ridder, A. van de Stolpe, E. van den Akker, W. Verheagh, M.J.T. Reinders, Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics, Sci Rep. 6 (2016) 24949. doi:10.1038/srep24949.
[64] C.R. Planey, O. Gevaert, CoINcIDE: A framework for discovery of patient subtypes across multiple datasets, Genome Med. 8 (2016) 27. doi:10.1186/s13073-016-0281-4.
[65] F. Vitali, L.D. Cohen, A. Demartini, A. Amato, V. Eterno, A. Zambelli, R. Bellazzi, A Network-Based Data Integration Approach to Support Drug Repurposing and Multi-Target Therapies in Triple Negative Breast Cancer, PLOS ONE. 11 (2016) e0162407. doi:10.1371/journal.pone.0162407.
[66] S. Marini, I. Limongelli, E. Rizzo, A. Malovini, E. Errichiello, A. Vetro, T. Da, O. Zuffardi, R. Bellazzi, A Data Fusion Approach to Enhance Association Study in Epilepsy, PLOS ONE. 11 (2016) e0164940. doi:10.1371/journal.pone.0164940.
[67] S.G. Ge, J. Xia, W. Sha, C.H. Zheng, Cancer Subtype Discovery Based on Integrative Model of Multigenomic Data, IEEE/ACM Transactions on Computational Biology and Bioinformatics. 14 (2017) 1115–1121. doi:10.1109/TCBB.2016.2621769.

[68] A. Ahmad, H. Frohlich, Towards Clinically More Relevant Dissection of Patient Heterogeneity via Survival based Bayesian Clustering., Bioinformatics. (2017). doi:10.1093/bioinformatics/btx464.
[69] D. Toddenroth, T. Ganslandt, C. Drescher, T. Weith, H.-U. Prokosch, J. Schuettler, T. Muenster, Algorithmic Summaries of Perioperative Blood Pressure Fluctuations, Stud Health Technol Inform. 228 (2016) 532–536.
[70] B.D. Lehmann, J.A. Bauer, X. Chen, M.E. Sanders, A.B. Chakravarthy, Y. Shyr, J.A. Pietenpol, Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies, J. Clin. Invest. 121 (2011) 2750–2767. doi:10.1172/JCI45014.
[71] Z.C. Wang, N.J. Birkbak, A.C. Culhane, R. Drapkin, A. Fatima, R. Tian, M. Schwede, K. Alsop, K.E. Daniels, H. Piao, J. Liu, D. Etemadmoghadam, A. Miron, H.B. Salvesen, G. Mitchell, A. DeFazio, J. Quackenbush, R.S. Berkowitz, J.D. Iglehart, D.D.L. Bowtell, Australian Ovarian Cancer Study Group, U.A. Matulonis, Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome, Clin. Cancer Res. 18 (2012) 5806–5815. doi:10.1158/1078-0432.CCR-12-0857.
[72] J.-H. Kwon, N. Lee, J.Y. Park, Y.S. Yu, J.P. Kim, J.H. Shin, D.-S. Kim, J.W. Joh, D.S. Kim, K.Y. Choi, K.-J. Kang, G. Kim, Y.H. Moon, H.J. Wang, Actionable gene expression-based patient stratification for molecular targeted therapy in hepatocellular carcinoma, PLoS ONE. 8 (2013) e64260. doi:10.1371/journal.pone.0064260.
[73] C. Pak, N.S. Callander, E.W.K. Young, B. Titz, K. Kim, S. Saha, K. Chng, F. Asimakopoulos, D.J. Beebe, S. Miyamoto, MicroC(3): an ex vivo microfluidic cis-coculture assay to test chemosensitivity and resistance of patient multiple myeloma cells, Integr Biol (Camb). 7 (2015) 643–654. doi:10.1039/c5ib00071h.
[74] D.E. Faries, Y. Chen, I. Lipkovich, A. Zagar, X. Liu, R.L. Obenchain, Local control for identifying subgroups of interest in observational research: persistence of treatment for major depressive disorder, Int J Methods Psychiatr Res. 22 (2013) 185–194. doi:10.1002/mpr.1390.
[75] M.B. Stein, A. Keshaviah, S.A. Haddad, M. Van Ameringen, N.M. Simon, M.H. Pollack, J.W. Smoller, Influence of RGS2 on sertraline treatment for social anxiety disorder, Neuropsychopharmacology. 39 (2014) 1340–1346. doi:10.1038/npp.2013.301.
[76] F.W.K. Cheung, M.Y.Y. Law, A novel conformity index for intensity modulated radiation therapy plan evaluation, Med Phys. 39 (2012) 5740–5756. doi:10.1118/1.4742848.
[77] X. Chai, M. van Herk, A. Betgen, M. Hulshof, A. Bel, Automatic bladder segmentation on CBCT for multiple plan ART of bladder cancer using a patient-specific bladder model, Phys Med Biol. 57 (2012) 3945–3962. doi:10.1088/0031-9155/57/12/3945.
[78] D.E. Henson, A.M. Schwartz, D. Chen, D. Wu, The clinical implications of integrating additional prognostic factors into the TNM, J Surg Oncol. 109 (2014) 391–394. doi:10.1002/jso.23525.
[79] G. Maccarrone, C. Ditzen, A. Yassouridis, C. Rewerts, M. Uhr, M. Uhlen, F. Holsboer, C.W. Turck, Psychiatric patient stratification using biosignatures based on cerebrospinal fluid protein expression clusters, J Psychiatr Res. 47 (2013) 1572–1580. doi:10.1016/j.jpsychires.2013.07.021.
[80] M. Deliu, D. Belgrave, M. Sperrin, I. Buchan, A. Custovic, Asthma phenotypes in childhood, Expert Rev Clin Immunol. (2016) 1–9. doi:10.1080/1744666X.2017.1257940.
[81] S.I. Rennard, N. Locantore, B. Delafont, R. Tal-Singer, E.K. Silverman, J. Vestbo, B.E. Miller, P. Bakke, B. Celli, P.M.A. Calverley, H. Coxson, C. Crim, L.D. Edwards, D.A. Lomas, W. MacNee, E.F.M. Wouters, J.C. Yates, I. Coca, A. Agustí, Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints, Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis, Ann Am Thorac Soc. 12 (2015) 303–312. doi:10.1513/AnnalsATS.201403-125OC.

[82] A.V. Zinchuk, M.J. Gentry, J. Concato, H.K. Yaggi, Phenotypes in obstructive sleep apnea: A definition, examples and evolution of approaches, Sleep Med Rev. (2016). doi:10.1016/j.smrv.2016.10.002.
[83] H. Bolouri, L.P. Zhao, E.C. Holland, Big data visualization identifies the multidimensional molecular landscape of human gliomas, Proc. Natl. Acad. Sci. U.S.A. 113 (2016) 5394–5399. doi:10.1073/pnas.1601591113.
[84] S.-A. Brown, Patient Similarity: Emerging Concepts in Systems and Precision Medicine, Front Physiol. 7 (2016). doi:10.3389/fphys.2016.00561.
[85] R.A. Greenes, ed., Clinical Decision Support: the road to a broad adoption, Academic Press, Oxford, 2014.
[86] WHO, Disease burden - Estimates for 2000–2015, (n.d.). http://www.who.int/healthinfo/global_burden_disease/estimates/en/ (accessed January 31, 2018).
[87] Global Burden of Disease Cancer Collaboration, Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study, JAMA Oncology. 3 (2017) 524–548. doi:10.1001/jamaoncol.2016.5688.
[88] C.D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030, PLoS Med. 3 (2006) e442. doi:10.1371/journal.pmed.0030442.
[89] UNICEF, World Health Organization, Diarrhoea: why children are still dying and what can be done, UNICEF, World Health Organization, New York, 2009.
[90] B.M. Welch, K. Kawamoto, Clinical decision support for genetically guided personalized medicine: a systematic review, Journal of the American Medical Informatics Association. 20 (2013) 388–400. doi:10.1136/amiajnl-2012-000892.
[91] J.D. Tenenbaum, P. Avillach, M. Benham-Hutchins, M.K. Breitenstein, E.L. Crowgey, M.A. Hoffman, X. Jiang, S. Madhavan, J.E. Mattison, R. Nagarajan, B. Ray, D. Shin, S. Visweswaran, Z. Zhao, R.R. Freimuth, An informatics research agenda to support precision medicine: seven key areas, J Am Med Inform Assoc. 23 (2016) 791–795. doi:10.1093/jamia/ocv213.
[92] A. Sharafoddini, J.A. Dubin, J. Lee, Patient Similarity in Prediction Models Based on Health Data: A Scoping Review, JMIR Med Inform. 5 (2017) e7. doi:10.2196/medinform.6730.
[93] W. Astle, D.J. Balding, Population Structure and Cryptic Relatedness in Genetic Association Studies, Statistical Science. 24 (2009) 451–471. doi:10.1214/09-STS307.
[94] J. Yu, G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, E.S. Buckler, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, Nat Genet. 38 (2006) 203–208. doi:10.1038/ng1702.
[95] Z. Zhang, E. Ersoz, C.-Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, J.M. Ordovas, E.S. Buckler, Mixed linear model approach adapted for genome-wide association studies, Nat Genet. 42 (2010) 355–360. doi:10.1038/ng.546.
[96] A.L. Price, N.A. Zaitlen, D. Reich, N. Patterson, New approaches to population stratification in genome-wide association studies, Nat Rev Genet. 11 (2010) 459–463. doi:10.1038/nrg2813.
[97] A. Korte, A. Farlow, The advantages and limitations of trait analysis with GWAS: a review, Plant Methods. 9 (2013) 29. doi:10.1186/1746-4811-9-29.
[98] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, Sci Rep. 6 (2016) 26094. doi:10.1038/srep26094.
[99] M. Liang, Z. Li, T. Chen, J. Zeng, Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach, IEEE/ACM Trans Comput Biol Bioinform. 12 (2015) 928–937. doi:10.1109/TCBB.2014.2377729.

[100] A.K. Goel, B. Diaz-Agudo, What's Hot in Case-Based Reasoning., in: AAAI, 2017: pp. 5067–5069. http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/15041/14020 (accessed July 11, 2017).
[101] D. Gu, C. Liang, H. Zhao, A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis, Artif Intell Med. 77 (2017) 31–47. doi:10.1016/j.artmed.2017.02.003.
[102] J. Ramos-González, D. López-Sánchez, J.A. Castellanos-Garzón, J.F. de Paz, J.M. Corchado, A CBR framework with gradient boosting based feature selection for lung cancer subtype classification, Comput. Biol. Med. 86 (2017) 98–106. doi:10.1016/j.compbiomed.2017.05.010.

Author Contributions Statement

E.P., S.M. and L.S. extracted and analyzed the papers and wrote the manuscript. R.B. originally proposed the topic of the article, supervised the study and contributed to the introduction and discussion sections. All authors reviewed the manuscript.

Additional information

Competing financial interests: the authors declare no competing financial interests.

Figure legends text

Figure 5 - Article selection flow diagram

Figure 6 – Number of articles classified according to the types of data used to compute patient similarity

Figure 7 – Number of articles classified according to the clinical domain

Figure 4 – Number of articles employing a novel approsach classified according to the data analysis methods employed for similarity

Highlights

- Ever-growing size and availability of health-related data is challenging the broad definitions of guideline-defined patient groups
- Defining a patient similarity measure is essential to allow stratification of patients into clinically-meaningful subgroups
- The present review investigates the use of patient similarity as a tool to enable precision medicine

Graphical abstract