Models and Data Sources Used in Systems Medicine*

A Systematic Literature Review

M. Gietzelt; M. Löpprich; C. Karmen; P. Knaup; M. Ganzinger
Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany

Keywords
Systems medicine, systematic literature review, data sources, models, medical informatics

Summary

Background: Systems medicine is a new approach for the development and selection of treatment strategies for patients with complex diseases. It is often referred to as the application of systems biology methods for decision making in patient care. For systems medicine computer applications, many different data sources have to be integrated and included into models. This is a challenging task for Medical Informatics since the approach exceeds traditional systems like Electronic Health Records. To prioritize research activities for systems medicine applications, it is necessary to get an overview over modelling methods and data sources already used in this field.

Objectives: We performed a systematic literature review with the objective to capture current use of 1) modelling methods and 2) data sources in systems medicine related research projects.

Methods: We queried the MEDLINE and ScienceDirect databases for papers associated with the search term systems medicine and related terms. Papers were screened and assessed in full text in a two-step process according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement guidelines.

Results: The queries returned 698 articles of which 34 papers were finally included into the study. A multitude of modelling approaches such as machine learning and network analysis was identified and classified. Since these approaches are also used in other domains, no methods specific for systems medicine could be identified. Omics data are the most widely used data types followed by clinical data. Most studies only include a rather limited number of data sources.

Conclusions: Currently, many different modelling approaches are used in systems medicine. Thus, highly flexible modular solutions are necessary for systems medicine clinical applications. However, the number of data sources included into the models is limited and most projects currently focus on prognosis. To leverage the potential of systems medicine further, it will be necessary to focus on treatment strategies for patients and consider a broader range of data.

Correspondence to:
Matthias Ganzinger
Heidelberg University
Institute of Medical Biometry and Informatics
Im Neuenheimer Feld 130:3
69120 Heidelberg
Germany
E-mail: matthias.ganzinger@med.uni-heidelberg.de

1. Introduction

In recent years, systems biology was highly successful in providing dedicated models for cellular processes and organisms. These models are a valuable basis for a better understanding of diseases with often very complex pathophysiologicals. However, due to the origin in systems biology, such models often focus on the description of disease-specific mechanisms, but rarely on the immediate treatment of patients. Thus, it is necessary to translate the results of systems biology into clinical practice in order to improve prospects of patients with severe diseases.

Systems medicine is such an approach for novel ways of treating complex diseases [1–4]. In contrast to systems biology, systems medicine strives for an individualized prognosis and treatment of patients based on heterogeneous data including omics data as well as phenotype data and individual preferences. Therefore, patient care with systems medicine is only possible with the help of a new kind of information technology (IT). Systems medicine IT will have to integrate all available data and apply disease models on individual patient data to provide decision support. As such, it combines features of Electronic Health Record (EHR) systems and clinical Decision Support Systems (DSS). However, coverage of traditional EHR data sets will not be sufficient. Data will include the whole spectrum of omics data, but also cover aspects like patient lifetime environment, lifestyle, routine clinical data, and others. Further, available data types will grow over time and IT systems will have to be flexible in including new data.

Designing a systems medicine application is a challenging task for Medical In-
formatics, since many data sources and models have to be considered and new algorithms as well as decision support strategies have to be implemented. This complexity suggests a modular, generic IT-architecture and an iterative development process. To select well-proven methods and data types for these tasks, it is important to prioritize models and data sources to be integrated into a systems medicine application. To get an overview of current models and data already used in systems medicine projects, we performed a systematic literature review. Our main research questions were:

1. Which modelling methods are used in systems medicine?
2. What kind of data and data sources are used in systems medicine?

These questions cover mathematical or statistical methods as well as any other approaches used for modelling. The answers to these questions are important for future systems medicine projects since they can provide guidance on the strengths and limitations of the methods. Further, systems medicine is also driven by huge amounts of data that become available both from the perspective of disease populations and individual patients. Thus, it was never before possible to describe patients in terms of their genotypes and phenotypes as comprehensive as now.

2. Methods

We performed a systematic literature review that complies with the guidelines outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [5]. PRISMA is a comprehensive evidence-based approach to support systematic reviews and meta-analyses. The PRISMA checklist for this review is available as Supplementary Online File.

2.1 Search Strategy

We conducted the literature search in the databases MEDLINE (via PubMed) and ScienceDirect. The query was developed in a query definition process resulting in the key words systems medicine and model. Target fields of the data sets in the literature databases were title and abstract. When developing the query, we also tested more constraining alternatives with the aim of filtering data sources as well. However, the term data source is often not present in the papers. Thus, a list of terms for specific data sources would have been necessary. Since such a list hardly can be complete, we decided not to include it into the query to avoid bias. We complemented our search strategy by adding the plural forms of the key words. Since personalized medicine is often used as a synonym for systems medicine in the domain of clinical research, we also added this term to the query. The query was translated into the query syntax of the selected databases as shown in Online Appendix Table 1. Due to the nature of our research questions, no specific measures against publication bias were implemented. Potential impact of this decision to our results is discussed in the paragraph "Limitations of the Review".

2.2 Inclusion Criteria

All papers retrieved through database queries were checked against a set of inclusion criteria. We included research reports covering disease models based on more than one different data sources. Studies using only one data source were excluded because in our opinion a comprehensive systems-oriented view cannot be achieved from the perspective of only one data source. In addition, studies included had to provide output that is the basis for a therapy decision. The rationale for this criterion is that systems medicine has the purpose of helping patients directly and not just describing the disease. The first database search was performed in May 2014. Thus, only articles published before May 2014 were considered in the systematic review.

Book chapters, editorials and letters to the editors were excluded from further investigation. Articles written in languages other than English or German were excluded as well.

2.3 Article Selection

In the next step, we selected articles based on their title and abstract according to the inclusion criteria. When abstracts were unavailable, full text articles were retrieved.

To test consensus among our four reviewers concerning the inclusion criteria, all reviewers assessed the same sample of approximately 10% of all articles independently. Articles with differing judgement were discussed during a meeting of all reviewers until consensus was reached. This helped to ensure that all reviewers applied the same understanding of inclusion criteria during the selection process.

Articles were equally split among four reviewers and each article was assigned to two reviewers initially. Reviewers categorized each article as ‘inclusion’, ‘exclusion’, or ‘unclear’. If the two reviewers disagreed on inclusion or in case of uncertainty, a third reviewer was consulted for an additional assessment of the article. If the third reviewer did not match the previous ones (an article was rated as inclusion by the first reviewer, unclear by the second reviewer and exclusion by the third, e.g.), a fourth reviewer finally decided on inclusion or exclusion of the article in question.

2.4 Data Extraction and Synthesis

For full-text analysis, the articles included were coded according to a deductively and iteratively devised coding system, using the software package MAXQDA for qualitative content analysis [6–8]. This software belongs to the category of computer-assisted qualitative data analysis software (CAQDAS). Alternative CAQDAS solutions with similar functionality are for example ATLAS.ti [9] or NVivo [10], but the licensing conditions of MAXQDA suited our requirements best. In contrast to manual approaches, a literature review with the help of CAQDAS is much more transparent and reproducible since codings can easily be traced back to text passages in

* A comprehensive overview and reviews on CAQDAS are maintained by the CAQDAS networking project: http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/ (last accessed 2015–12–23)
the original papers. We used MAXQDA for tagging relevant information in the articles with codes for later analysis. The papers were loaded into MAXQDA in Portable Document Exchange (PDF) format and coded by the four reviewers (a screenshot of MAXQDA is shown in Online Appendix Figure 1). Starting with an article with high degree of consensus among reviewers, relevant text passages were coded by open codes representing the text as closely as possible. Throughout reviewing of the full articles, the coding scheme was continuously extended. Synonymous codes were merged, combined and clustered into higher-ranking categories.

3. Results

3.1 General Results

The queries resulted in 609 articles from MEDLINE and 171 from ScienceDirect. After removing duplicates, 698 articles remained. The initially assigned reviewers agreed on inclusion and exclusion in 78.9 % (551 abstracts) of all abstracts retrieved. Online Appendix Table 2 shows the levels of agreement among the reviewers.

After the first round, 147 abstracts (21.1%) with conflicting review results were screened by a third, independent reviewer, who was only allowed to choose between including and excluding the abstract. In the second round, 39 additional abstracts were included, 96 were excluded. However, 12 articles (1.72%) rated in all three possible categories remained. In the final round, a fourth independent reviewer decided to include six of the remaining abstracts and to exclude another six. In total, 73 articles were included and 625 excluded. Main reasons for exclusion were that only one data type was used for analysis or the lacking of therapy indication based on the data. Another ten articles (1.43%) were excluded, because they were written in a language other than English or German. Online Appendix Figure 2 shows the PRISMA chart of the overall review process [5]. A Supplementary Online File lists all articles excluded in this phase together with the reasons for exclusion.

Table 1 Overview of modelling methods

<table>
<thead>
<tr>
<th>Category of Modelling</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning</td>
<td>26</td>
</tr>
<tr>
<td>Time series</td>
<td>3</td>
</tr>
<tr>
<td>Statistical tests/measurements</td>
<td>6</td>
</tr>
<tr>
<td>Networks</td>
<td>27</td>
</tr>
<tr>
<td>Others (e.g. Ordinary differential equation ODE)</td>
<td>21</td>
</tr>
</tbody>
</table>

3.2 Final Coding Scheme and Categories

In an iterative process the coding scheme and categories were derived from the articles included using the MAXQDA software. In a final step, the categories were revised, classified and clustered into new categories. The six main categories are 1) aim, 2) conclusion, 3) data source and category, 4) definition, 5) diseases and 6) modelling methods. Online Appendix Table 2 shows a complete export of all levels of the final coding scheme.

During coding process, the categories determined for category 5 – diseases indicated for the use of the International Statistical Classification of Diseases and Related Health Problems, 10th Revision, (ICD-10) as standard terminology instead of our own classification [11]. All articles previously coded were reviewed again and recoded according to ICD-10 codes.

3.3 Full Text Analysis

For all articles in this phase, we performed a full-text analysis. As a result, another 39 articles were excluded. These articles are listed in a Supplementary Online File together with the reasons for exclusion. In total, 34 articles were included for answering the research questions and coded according to the principles of qualitative content analysis [12–44]. The complete list of papers included for this review is shown in Online Appendix Table 3.

3.4 Modelling Methods in Systems Medicine

The modelling methods currently used in systems medicine are very heterogeneous and it was difficult to organize them in a mono-hierarchical structure. We established the categories machine learning methods, time series, statistical tests or measurements, network analysis methods, and others. However, this is only one way to classify these methods, because methods can be assigned to multiple categories.

Machine learning based clustering and other classification methods are used very often, but also network analysis methods and ordinary differential equations. Time series analyses, regressions and classical statistical methods like correlation and association measurements are applied less often. Bayes classification, clustering and tree induction methods were identified as representatives of machine learning methods. In Online Appendix Table 4 we summarize the methods found in the literature review.

Only 13 articles report about the tools used for analysing data. Examples are common tools like the statistical language R [45] or MATLAB [46], but also specialized software for particular data sources like Chenomx NMR Suite [47] for the analysis of data from Nuclear Magnetic Resonance Spectroscopy devices, or ICTNet [48] for the integration, visualization and analysis of genome-wide biological networks. In Online Appendix Table 5 we present a summary of the tools used in the studies included into this review.

3.5 Data Sources and Data Categories

We found 24 statements describing the origin of the data processed in systems medicine projects. These statements were divided into 11 categories of data sources as shown in Online Appendix Table 2. The most frequently mentioned data sources (17.9% each) are electronic medical records and public databases, such as Entrez Gene [49]. More than one-third (35.7%) of the reported data sources relate to the group of literature, population data, and personal medical history. Data from biobanks and drug-response-profiles each were used in 7.1% of the studies.

We divided data categories in systems medicine into three main categories: personal data about the subject (e.g. demo-
graphic data, behavioural data), clinical data (e.g., laboratory data, imaging data, examination data), and omics data. Most articles (48.4%) mention the use of omics data, especially genomics and molecular data (27.7%). About one-third (29.4%) incorporate clinical data, mostly laboratory data or data from examinations. Another 22.4% are working with subject data, which usually are limited to demographic or behavioural data. ►Online Appendix Table 6 shows the data types that were found during the review. A matrix of concurrently used data types within the same article is shown in ►Online Appendix Figure 3. Genomic data is the most frequently used data subcategory (21.4%). It is mostly used together with the subject’s demographic data (in nine cases) but also frequently together with other omics data, clinical examination data (seven cases), and laboratory data (five cases).

### 3.6 Definition of Systems Medicine

For better understanding how the term *systems medicine* is used, we provide an overview of the definitions we found in our analysis. Although we were mainly interested in articles on systems medicine, we retrieved many articles that have rather a strong focus on *systems biology*. In this section, we try to identify the similarities, differences and intersections between these terms.

Alfredo et al. consider systems biology the “analysis and reconstruction of biological networks at various levels of organization: from molecules, cells and tissues to organs, organisms, and groups of interacting organisms” [13]. Two main approaches can be distinguished: the “omics” approach based on data analysis, and the mathematical modelling approach [37]. In contrast, for Cardinal-Fernández et al. systems medicine is when the “concept of Systems Biology is applied to the field of health sciences” [19]. Both are “practically identical concepts, except [systems biology] is general and [systems medicine] is particularly focused on medicine” [19]. Vandamme et al. state, systems biology has the “potential to make a more translational impact in the arena of medical science” [50] while systems medicine is supposed to “aid physicians in handling more complex data in their day-to-day practice” [33]. For Wang et al. this can be done, for example, by detecting and stratifying various pathological conditions with the help of “clinically detectable molecular fingerprints resulting from disease-perturbed biological networks” [44], as discovered in the domain of systems biology. In the opinion of Tian et al., systems medicine is the idea that “disease arises as a consequence of the disease-perturbation (genetic and/or environmental) of one or more biological networks in the relevant organ” [3]. For Nibbe and Cardinal-Fernández the main goal for systems medicine is to “provide predictive models of the pathophysiology of complex diseases as well as define healthy states” [51] and to allow “new associations to be established between biological functions and special human diseases or conditions” [19].

### 3.7 Diseases Covered

To get an overview which diseases are covered by systems medicine approaches, we also coded the articles included for the respective diseases. For this coding we used the ICD-10 diagnosis codes. The same level of detail as described in the article was applied to the coding. The articles included led to 124 diagnosis codes with three digits like A00. In articles, where the three digit syntax could not be applied, the diagnosis codes were combined into blocks (e.g. A00–A09) and the parent category was coded.

Summarizing the top level of the ICD-10 classifications system, 71 occurrences of diagnosis codes were distributed across most of the 22 predefined ICD-10 chapters with a mean of 2.3 chapters addressed per paper and a maximum of ten. ►Online Appendix Table 8 shows the frequency of ICD-10 chapters as they occurred in the papers included. The most frequently mentioned code blocks were C00–D48 Neoplasms and I00–I99 Diseases of the circulatory system (both referenced by 11 papers).

### 3.8 Aims of Studies

The aims of systems medicine studies described in the articles included differ. Goals include, among others, prediction of diseases, cost reduction, decision support, and biorepositories. However, many studies aim to train models for decision support as data and knowledge representation, and to evaluate models in terms of evaluation and therapy outcome prediction. ►Online Appendix Table 9 presents a synthesis of the aims of systems medicine found in the literature.

### 4. Discussion

#### 4.1 Principal Findings

The literature review showed that a multitude of research activities is published with reference to systems medicine: About 700 papers were returned by the initial data-

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Frequency</th>
<th>Percentage</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic medical record</td>
<td>5</td>
<td>17.9</td>
<td>[12, 18, 23, 25, 34]</td>
</tr>
<tr>
<td>Public database</td>
<td>5</td>
<td>17.9</td>
<td>[15, 25, 29, 33, 42]</td>
</tr>
<tr>
<td>Literature</td>
<td>4</td>
<td>14.3</td>
<td>[12, 25, 27, 31]</td>
</tr>
<tr>
<td>Population data</td>
<td>3</td>
<td>10.7</td>
<td>[12, 18, 23]</td>
</tr>
<tr>
<td>Personal medical history</td>
<td>3</td>
<td>10.7</td>
<td>[12, 13, 31]</td>
</tr>
<tr>
<td>Biobank</td>
<td>2</td>
<td>7.1</td>
<td>[17, 34]</td>
</tr>
<tr>
<td>Drug-response-profile</td>
<td>2</td>
<td>7.1</td>
<td>[31, 33]</td>
</tr>
<tr>
<td>Actual drug prescriptions</td>
<td>1</td>
<td>3.6</td>
<td>[13]</td>
</tr>
<tr>
<td>Decision support system</td>
<td>1</td>
<td>3.6</td>
<td>[26]</td>
</tr>
<tr>
<td>Derivation study</td>
<td>1</td>
<td>3.6</td>
<td>[36]</td>
</tr>
<tr>
<td>Use of medical devices</td>
<td>1</td>
<td>3.6</td>
<td>[13]</td>
</tr>
</tbody>
</table>

Table 2 Overview of data sources described in the papers included
base query. However, a deeper analysis of these papers’ nature revealed that the majority of articles does not report on research where systems medicine has a direct impact on patient treatment. Many papers had to be excluded because they only provide information on a very specific aspect of the disease, which then is reported to be relevant for systems medicine later on.

The primary aim of this review was to find methods currently used in systems medicine. It turned out that there is a great variety of methods, with a focus on network analysis and machine learning methods. The modelling methods we found clearly show the ancestry from systems biology where models describing certain effects in biological systems have a long tradition. The studies we found often do not use these methods in a way directly affecting clinical practice. It will be interesting to observe if such systems biology methods will be used in decision support for patient care in the future. Similarly, the tools currently used in the studies are not very specific for the domain of systems medicine. For example, R and MATLAB are generic tools for statistical analysis and calculation; other tools like Microsoft Access are used for data management in general. The lack of specific systems medicine applications also affects literature research for such software from a Medical Informatics perspective: Our study has shown that it is hardly possible to formulate a precise query that does not include many irrelevant articles when searching scientific literature for systems medicine applications. It will be interesting to observe the future development of specialized systems medicine applications and its reflection in Medical Informatics literature.

A basic concept of systems medicine is using heterogeneous data for an individual treatment decision. Many studies in the context of systems medicine make use of omics data. However, numerous papers retrieved during the initial query had to be excluded, because only one data source was considered. Even the papers included hardly report on the use of more than two different data types. In these studies, mostly omics data are combined with demographic or clinical data, or several types of omics data are included. The prevalence of omics data could be explained by their major role in systems biology.

For a truly systems oriented approach it seems to be important to get a comprehensive view on the subject, and to consider as many aspects as possible. Here, it might be helpful to look back to the beginnings of systems theory. Bertalanffy wrote as early as 1950 in his paper An outline of general system theory: "Virchow's programme of 'cellular pathology,' claiming to resolve disease into functional disturbances of cells, is to be supplemented by the consideration of the organism-as-a-whole, as it appears clearly in such fields as theory of human constitutions, endocrinology, physical medicine and psychotherapy" [52]. In the era of big data this could mean to include multiple, if not all available data on patients and their diseases into systems medicine [53, 54]. In our review, we did not yet find projects that add data on the patients’ environment, live style, nutrition, and others as data source. These data are hard to measure since they cover a very long time span and lead to a huge amount of data, especially if they are measured with a high resolution. They might play a key role in the future for fighting diseases with a long delay between the event causing the disease and first symptoms. The currently emerging quantified self movement might help to generate such data [55].

Systems medicine is not limited to certain medical fields. Our results show that only three ICD-10 chapters are currently not addressed by systems medicine projects. This is a very promising result, since the whole spectrum of human diseases including organic, mental, and behavioural aspects is considered almost completely. According to our results, the most important aims for systems medicine are currently to determine the patients’ prognosis and finding new treatment possibilities for patients. This finding is consistent with the findings on reported modelling methods. The modelling methods are mostly used to describe biological mechanisms, which can be used to estimate prognosis and find new targets for treatment.

However, the term systems medicine is often not clearly distinguished from the term systems biology. Both terms are based on the systems approach. In this paradigm, disease is the consequence of the disturbance of specific entities in the vast quantity of elements of a biological system. The point of view and granularity define the difference between systems biology and systems medicine. Systems biology is strongly focused on the biological level, such as cells, molecules, proteins and their interactions. On the other hand, systems medicine defines a complete medical condition as the system covered. The fine-grained molecular level resulting from systems biology can be incorporated as elements among others in the disease system.

4.2 Limitations of the Review

In this study, only papers with more than one different data source are included. This number was chosen because at least two sets of data are necessary to indicate a statistical relationship involving dependency and correlation. However, two sources might still be considered a low number for a systems approach. On the other hand, only papers explicitly stating data sources were included. There might be papers implicitly using multiple data sources without mentioning them in the text.

The coding system we developed also has some deficiencies: Some modelling methods could be classified for multiple categories of modelling. An alternative to our approach would be to allow multiple parent relationships in the classification system. However, this issue mostly affects the presentation of the results and not the findings of the review in general.

Like any systematic review, this study is susceptible to publication bias. Thus, there might be modelling methods or data sources that are used in systems medicine which were not captured in this review. This might be especially true for approaches that did not result in a working systems medicine application. Our research questions address properties of systems medicine in general, not in a specific medical context. In contrast, reports on systems medicine research are mostly on specific applications for a certain disease area. If such a specific approach fails, we cannot necessarily conclude that the model as well as the data sources are not useful for
systems medicine in general. Thus, the bias introduced by not reporting such a specific result should not substantially harm the results of our systematic review.

5. Conclusions

Our review provides a current snapshot on the modelling methods and data sources used for systems medicine. While systems medicine is still a young discipline, we found many promising activities in its context. Many different modelling approaches are currently used to establish disease models in all areas of medical practice. However, to further expand the systems medicine approach, more data sources will have to be considered to pave the way to optimize treatment support for patients.

Acknowledgment

This literature review was done within the systems medicine project “Clinically-applicable, omics-based assessment of survival, side effects, and targets in multiple myeloma” (CLIOMICS). The project is funded by the German Federal Ministry of Education and Research (BMBF, grant id: 01ZX1309A) as part of the e:Med initiative.

References


Methods facebook community: www.facebook.com/MethodsInfMed