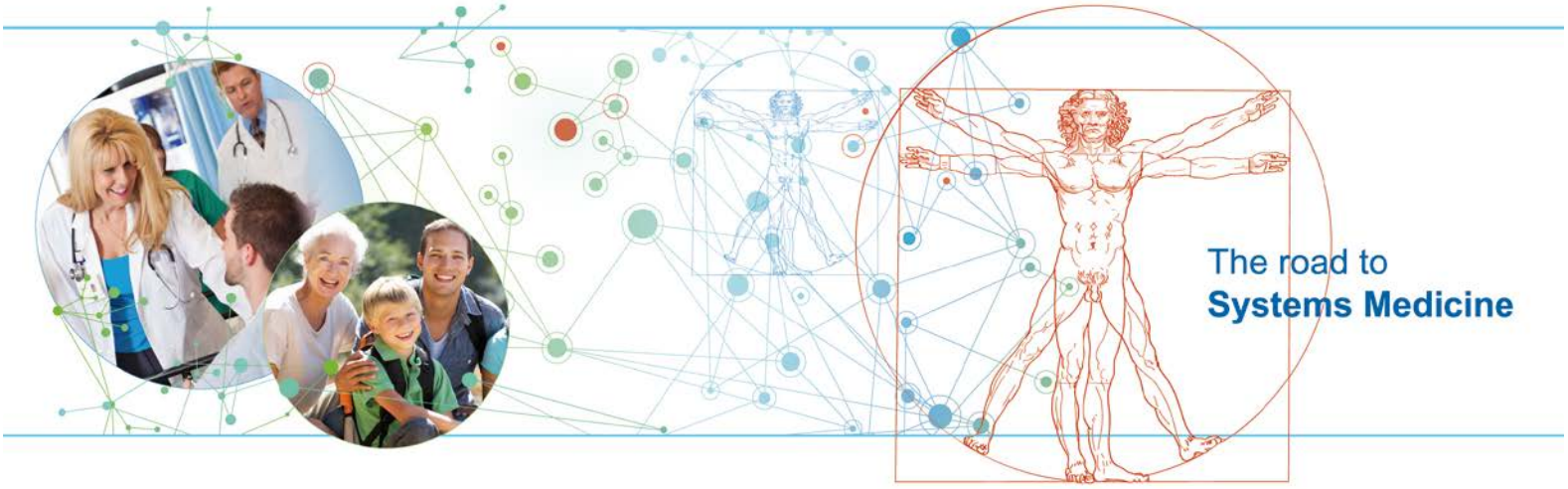
CASYM EUROPE

Coordinating Action Systems Medicine
Implementation of Systems Medicine across Europe

SEVENTH FRAMEWORK PROGRAMME



Requirements for supporting standardization and data sharing in Systems Medicine

CASyM report

April 2017

IMPRINT

Publisher

CASyM administrative office
Project Management Jülich, Forschungszentrum Jülich GmbH
m.kirschner@fz-juelich.de

Authors

Johannes Schuchhardt, Jesper Tegnér, Olaf Wolkenhauer

Date

April 2017

Contact information

Jesper Tegnér, Karolinska Institutet, Sweden
jesper.tegner@ki.se

References

CASyM report: Requirements for supporting standardization and data sharing in Systems Medicine, April 2017.

TABLE OF CONTENT

REQUIREMENTS FOR SUPPORTING STANDARDIZATION AND DATA SHARING IN SYSTEMS MEDICINE

Introduction

Here we provide some reflections on standardization and data sharing, and doing so, we make a few remarks upon topics, which are (implicitly) entailed when considering sharing and standards. These reflections are not only a product of discussions within the context of CASyM, but also fan out into other past or ongoing projects where we are involved as investigators.

First a general comment: It is well known and it has been substantially documented that the landscape is severely fragmented with regard to what are the different issues, different needs, and who are the different stakeholders/organizations involved or dependent upon the landscape. Discussions can easily become confusing if these aspects are not clearly identified or distinguished.

Hence, due to this situation it is not really feasible to summarize it all in one sentence capture what to do. An approximation of what would be a desirable outcome could well be "*disrupt the fragmentation*", which is not identical to the challenge on how to actually accomplish that. We we'll return to that in the end.

In our reflections below we first comment on some of the different topics involved, next we briefly survey the different needs with special reference to the different sub-communities of systems and computational research and applications. Third, we relate these points to different stakeholders with an interest in these topics or in need of a resolution of these topics. Finally, we close by considering how these reflections play into the subtopics as outlined in the CASyM roadmap.

Conceptual distinctions differentiating various challenges, bottle-necks, and opportunities regarding data and analysis

Size and complexity

In our view there is a perception that the big challenge is the big data in life science. This notion generates discussions on data-clouds, and how to share huge volumes of data. Furthermore, this leads to the issue of where to analyse these large data, since they are not readily moved. Hence, numerous discussions target how to situate the analysis machinery close to the data (in the cloud) to make this endeavour feasible. Our view is that these discussions are only in part relevant, and less tangential when considering full length DNA sequence data. Yet, in our shared experience we find that the heterogeneity of the data is much more tenacious challenge. As a rule, such data (when processed individually) is not that large, and can therefore readily be shared and moved around from a technical standpoint, notwithstanding legal/ethical issues here, which again are different. Hence, it´s not the volume, but rather the complexity of the different data-types, ranging from multiple molecular data, images, sensors, and up to health records. The challenge is rather how to harness the value from these data, a task which is different and to be specified depending on the particular

question. Evidently, standards are required to make sharing and integration of different data-types possible. This is particularly important when it comes to healthcare data, typically generated in distributed locations, different hospitals, institutes, regions and countries, where IT infrastructures, formal regulations, and soft practices for data handling do differ widely in our experience. This shift in perception on the nature of the core challenges also implies that *topics such as data standards and quality in particular become more pressing than the (data) sharing part* (excluding a discussion of health care data). With suboptimal standards and quality there is obviously less point in sharing.

Access & Sharing

Overall, access to data is good as long as we target molecular data of different kinds. One area, which we have less experience with, is molecular imaging, and to the best of our knowledge, these are indeed large data, requiring high levels of domain expertise to be useful as it stands now. However, the need for access and sharing appears to be less since it is not clear how to integrate such image information with molecular data, i.e. the point of complexity above. Worse is the case for healthcare data, including images (MRI, PET, x-rays). This access problem is a major roadblock towards the application of systems and computational tools bearing to the challenge of personalized and precision medicine. Currently, despite "good-intention-policy efforts", we are in practice left with ad-hoc solutions between collaborating teams/investigators. In conclusion, *sharing health care data is a major transnational urgent topic where we as of now, in our experience, have no good operational solutions in sight.*

Quality, Standards & Harmonization

We find the area of molecular data pretty matured in this aspect. For most of the individual data-types there are extensive open scripts and documentation on QC metric for such data. When considering clinical data, health care records, the field is in our hands very nascent. Metric for quality is not there, data sets are incomplete, not easily comparable when considering different cohorts for example. How to standardize or ensure that different protocols produce comparable data of "similar" quality? Simple things like "smoking" may mean different things in different data-sets/cohorts. Hence, the topic of harmonization between different clinical variables is a major challenge. Currently, this translates into manual time-consuming error-prone curation of data. Thus, the potential leverage of using large-scale data either on a clinical cohort in conjunction with molecular data or cross cohort analysis is in practice untested. In contrast, almost any service running over Internet or mobile devices hinges on the value creation of working with large, accessible data, of different types. Even for common diseases, when it comes to an individual patient, s/he is a 'rare case'. The genetic variations in humans are vast and the role of exogenous variables, including life style and environment, make it very difficult to compare cases. Such comparisons of an individual with a (sub) population are however the basis for diagnosis, prognosis and therapy and therefore, it is essential to pool experience in diagnosis and therapy across hospitals, regions and countries. *We find this dimension together with the access/sharing problem to constitute the major roadblock on a technical level for personalized/precision medicine, as we urgently need those multi-dimensional clinical phenotypes in conjunction with other data-types*. Our strong recommendation is that this is where concerted efforts are needed.

Different communities – different needs

Clearly different communities in systems and computational investigations active in research and industry have different needs with regard to the topics discussed above. Here we would like to distinguish between (a) Bioinformatics (pipeline/workflow analysis mainly focused on molecular data-types), (b) Computational Biology (algorithms, modelling, and simulations focused on spatio-temporal processes), (c) Systems Biology (integrated computational and experimental work targeting chiefly mechanistic understanding of a systems or a process), (d) Systems Medicine (medical focus, prediction, stratification, drug response, and to some degree disease understanding). *Using this, obviously simplified, blueprint of communities we find that it is clearly the Systems Medicine community who suffers most of the shortcomings of the limited access, sharing, and quality, of rich clinical phenotypic information/data*. This is not to say that all aspects of quality, standards, sharing etc. are solved with regard to the other communities, but our point is that, relatively speaking, the challenges the current data-situation in medicine hampers Systems Medicine the most, and personalized and precision medicine by extension.

The landscape and current situation of stakeholders

Given the reflections above we find that research communities dependent on the clinical data are those who suffer most. This is valid both for research, biotech, and pharmaceutical industry. The solution in industry has instead been to build internal databases of such data in narrow highly controlled collaborations. Again making these data in practice not accessible, since such data -in such an environment- clearly has a potential commercial value. Hence, the problem is not solved.

Here we would like to make the point that the current situation is not beneficial for the hospitals themselves, since they cannot easily cross-link their data, nor harness the potential value from such data. Thus, it should be in the interest of health care organizations to truly address this problem. However, this is complicated by legal/ethical topics, which also make the argument for administrators and IT staff within healthcare organizations to remain passive in this very note. The analysis of data in clinical settings across Europe falls short of a service to those who would benefit from data-driven diagnosis, prognosis and therapy. It is obvious that technological and methodological advances have been missed almost completely, for reasons that have nothing to do with science and healthcare.

In our experience we see two developments challenging this status quo. On the one hand, patients and citizens taking charge of their own data. This goes from recording and monitoring their bodies, activities, and mental states (sensors) and when doing so, being very open and sharing their data, e.g. PatientsLikeMe[1]. Secondly, private organizations, such as Mayo Clinic[2], have built a very data-driven systematic strategy to collect and annotate multiple levels of clinical and molecular information in their daily operation. Alternative, large IT companies are making inroads in this space, where Google´s very much debated access to hospital data in UK is one recent example. Once analyses can be conducted remotely, we may see a development where patients decide on their health by seeking help beyond their local doctors and clinics.

[1] **PatientsLikeMe** is a patient network and real-time research platform. Website: https://www.patientslikeme.com/

[2] Mayo Clinic is a nonprofit organization committed to clinical practice, education and research, providing expert, whole-person care to everyone who needs healing. Website: http://www.mayoclinic.org/

A slight speculative prediction is that change will not likely emerge within the established organizations such as the national health care systems and their hospitals.

Relevance for the CASyM roadmap

Computational biology, bioinformatics and systems medicine have developed tremendously along with technological developments to generate a wide range of data types. While the potential for these developments to improve the diagnosis, prognosis and therapy are obvious, we have noticed that the implementation of new approaches is hampered by problems that have largely structural and organizational reasons. In many cases, it seems, there is no doubt about benefits to patients with new approaches but these can nevertheless not be implemented in current legislative and financial structures. What this suggests is a need for a more patient-driven and political agenda to ensure patients can benefit from scientific and technological advances.

While the technological basis for the generation of data and their storage is relatively trivial from an IT perspective, the analysis of data however relies on the sharing and integration of data and this relies on standards. These standards have to be agreed upon by a large number of stakeholders and interest groups, which requires a substantial effort.

The reflections above are of relevance for several of the sections in the CASyM roadmap. In short and in accordance with the ten key areas of the CASyM roadmap (version 1 of June 2014) for a successful implementation of Systems Medicine we find that the challenge of *Patient Stratification* (roadmap key area 5) depends on the clinical phenotype data in order to be successful. However advanced our molecular and other technologies are and will further improve in the future, they need to be assessed and evaluated using their corresponding clinical traits/outcomes. Hence, we recommend that the topics of *Methodological and Technological development including modelling* (roadmap key area 2), *Data generation* (roadmap key area 3) and *Technological infrastructure* (roadmap key area 4) should be geared towards developing technical tools, systems, algorithms, that enables generation of high quality clinical data, which can be shared in a secure manner. Moreover, there is a need to develop algorithms and software that could facilitate detection of anomalies in large unstructured health care data sets (roadmap key area 2). For example, what is missing, can´t be readily compared, what is of lower or uneven quality, and how to “normalize and harmonize” data in a data-driven manner. All towards mitigating the current need for relying completely on manual curation. *This should be one of the high priorities in the era of personalized/precision medicine (roadmap key area 5) and here the Systems Medicine community could play a major role.*

Conclusion

Now, in concluding, it goes without saying that there is an urgent need to disrupt the fragmentation. It is also evident that communities targeting in the end high-end goals of precision and personalized medicine are those very investigators who suffer the most from the current balkanization. Systems approaches, holding an impressive toolkit under their west, are in practice restrained to unleash machine intelligence, very much in contrast to the revolutions taking place in the digital world of devices, which are monetized by the major companies in this space. The values lost for patients and healthcare should not only be quantified in dollars but rather in unnecessary suffering and loss of control of your own wellbeing.

ACKNOWLEDGEMENTS

The present report is part of the CASyM Work Package 3 – "The technological and methodological basis of systems medicine" and Task 3.2./D3.2: Identification of requirements that support standardization and data sharing/and workshop report.

CASyM is funded by the European Union, Seventh Framework Programme under the Health Cooperation Theme and Grant Agreement # 305033.

Steering Committee

The following officials, as part of the Scientific Steering Committee, are involved in the scientific coordination of CASyM:

Charles Auffray - European Institute for Systems Biology & Medicine - EISBM, France
Mikael Benson (Deputy Chair) - Linköping University Hospital, Sweden
Rob Diemel - The Netherlands Organisation for Health Research and Development, The Netherlands
David Harrison (Chair) - University of St. Andrews, United Kingdom
Walter Kolch - University College Dublin, Ireland
Johannes Mohr - Federal Ministry of Education and Research, Germany
Francis Lévi - Institut National de la Sante et de la Recherche Medicale, France
Damjana Rozman (Deputy Chair) - University of Ljubljana, Faculty of Medicine, Slovenia
Johannes Schuchhardt - MicroDiscovery GmbH, Germany
Olaf Wolkenhauer - Dept. of Systems Biology & Bioinformatics University of Rostock, Germany

Administrative Office (Coordination)

Marc Kirschner - Project Management Jülich, Forschungszentrum Jülich GmbH, Germany


CASYM
EUROPE
SEVENTH FRAMEWORK
PROGRAMME