IMI2 JU Scientific Committee recommendations regarding data infrastructure and integration

Summary

Data analytics in biomedicine is still at an early stage. The field is now developing the technology to handle and make interoperable large data sources, making possible the reuse of the data. The rational integration of heterogeneous data represents still a basic conceptual and technical challenge for the effective use large multi-dimensional data sources. The development of systems for the optimal allocating the data, software and computational resources necessary to solve biomedical questions is another key area of development.

Artificial intelligence/machine learning (AI/ML) and computational simulations are now developed at unprecedented levels. Improving and integrating them seems the most logical step towards the development of mechanistic causal models that can link the molecular and physiological levels, and provide solid grounds to the prediction of diseases and disease -associated interventions.

Finally, it is important not to forget that the final aim of all these technical developments should be to provide better health systems levelling the current social heterogeneities.

Data and methods

A first wave of collaborative IMI projects addressed data-related issues, including aspects of real world data, ontologies and FAIRification[1]. Operating on secure distributed data will require the development of an additional software layer on top of the data to solve simultaneously the data access, security and analysis issues. In parallel with data access, the systems will have to integrate the analytical capacity to exploit the data in response to medical questions. Furthermore, the systems should be transparent to the end-users (i.e. user will not need to intervene directly at each step of data discovery and access), possibly based on cryptography technology.

At the technical level, containerised software for the sensitive data handling and workflow enactment strategies across the *computing continuum* should be the basic components[2]. Workflow enactment and data protection technologies will need to be brought into a common framework to simplify the adoption of safe and responsible practices.

To be able to compute the large distributed complex datasets the adequate information technology (IT) systems scaling from edge devices to high-performance computing (HPC) will be required. The computing continuum concept implies the technology to analyse data with the right methods in the adequate computing environment: HPC centers, cloud and clusters, local computers to edge devices.

The essential IT component has been largely neglected in previous IMI projects, while part of this technology is developed in other areas (DG Connect) but rarely applied to biomedical data. In order to navigate the requirements posed by data protection and computational demands, technology will need to be developed to facilitate reliable and reproducible work orchestration across different computing environments from large HPC centres and distributed cloud systems to personal devices.

---

[1] IMI1 Open PHACTS and eTOX project and IMI2 FAIRPLUS and eTRANSAFE projects. Additionally, the **IMI BD4BO big data initiative** embraces a number of IMI2 projects: HARMONY; Big Data in Hematologic Malignancies, EHDEN; European Health Data and Evidence Network, BigData@Heart; Big Data in Heart Disease, PIONEER; Big Data in Prostate Cancer, and ROADMAP; Big Data in Alzheimer 's disease.

[2] The **Patient dossier** provides the conceptual framework for linking the questions formulated by the end users with the answers in large distributed data systems. (Users formulate abstract medical questions; the system will decompose the queries in smaller questions addressable with a combination of the adequate data, software and compute, and orchestrators, respecting the appropriate data management and access policies. The technical elements to build patient dossiers exist in the form of the current workflow technologies.)

Data and compute will only make sense if exploited by the new AI/ML based (Deep Neural Networks) and simulation methodologies, opening new opportunity for the development of analysis and decision support systems. Mechanistic simulations are the instrument to represent and exploit detailed knowledge on biomedical systems; a mechanistic knowledge that is essential to sustain progress on solid scientific basis. Simulations must be setup based on the most reliable set of assumptions and data, which requires carefully designed knowledge management strategies. Furthermore, simulations consume and produce benchmarking datasets. The development of appropriate surrogate (simulated) data sets, including avatars and digital twins[3], is a particularly attractive strategy in many areas of biomedicine.

Directly associated to the development of data access and analysis systems it will be fundamental for industry and society to have access to systems to guaranteeing the reproducibility and reusability of the computational methods. Benchmarking systems able to continuously assess at the technical and scientific level the performance of the systems, making the results of the benchmarking openly accessible will increase the confidence of users and regulators. Benchmarking systems, by providing data for training and testing, will stimulate development and foster industry development. Regarding, the application of AI/ML technologies in biomedicine, the issues of accountability are particularly important and regulatory required. A number of efforts are ongoing to develop systems under what is known as Explainable Artificial Intelligence[4].

The complexity of the new technical developments, even if offered to the end users in the appropriate form (adequate access systems), will require a sizeable effort to train and adapt the medical systems to exploit them, addressing issues of specialisation, internal compartmentalisation and regulations of the different European health systems.

Additionally, the efforts to build infrastructures to make medical data accessible and interoperable will have to be complemented by the development of the regulatory and commercial channels to make the exploitation of health data economically viable and socially acceptable. Data sharing constraints impose *in-situ* analyses where the data resides, demanding trained operators across institutions.

Recommendations

Two toy examples of data use in medical practice:

- *A clinician needs to check if medical image from her case matches previous cases in which a diagnosis could be based on.*

- *A hospital emergency system needs to assess rapidly the possible options to treat an accident, considering the patient's previous medical conditions.*

The system should allow them to: **(a)** formulate the questions in their own language, considering that the images or the medical reports on previous diagnosis might be located in more than one distant repository. Such a system will have to **(b)** localise the information in the distributed servers, and access it with the proper user credentials, **(c)** send the information under a secure protocol to the right computer with enough capacity and the right software (image comparison in one case, simulation of interventions in different medical conditions). The system will have to **(d)** summarise the information and give them the key reasons (e.g. features in the images and or critical points in the simulations) for the proposed recommendations.

To be able to solve these questions, and the many other similar ones continuously present in the medical practice, a number of key aspects in data organisation and access have to be solved:

[3] **Medical Digital Twin** as the final goal of the integration of data access, security, distributed computing and AI/ML and simulations methodologies. Even if not currently reachable, remains as a powerful paradigm of future medical support systems which arrival will be worth accelerating with specific medical case specific projects.

[4] The DARPA project Explainable Artificial Intelligence (XAI) https://www.darpa.mil/program/explainable-artificial-intelligence. European initiative AI-on-Demand initiative https://www.ai4eu.eu/ among others.

a) Clinicians should be able to formulate their medical questions in their own language. The technology to **build the "patient dossiers"** in response to the question will have to decompose the queries in smaller questions addressable with a combination of the adequate data, software and compute. To be useful data will have to be organised in standard, organised databases, with the proper labels (metadata) and associated medical information (EHRs). Data will have to be accessible by analytical workflows able to scan the database information and will have to include the legal accreditations following the General Data Protection Regulation (GDPR) and specific regulations.

1) Implement IT technology (database/software systems) integrating the data security (authentication and authorization[5]) and access the data required by the medical decision support systems.

2) Address the key issues of consistency and completeness of the medical data, in the context of the dynamic and heterogeneous (area and language specific) nature of health data.

3) Explore the systems using encrypted distributed data[6] to overcome data privacy issues.

b) Finding and accessing the required data. The data sources will be in many cases of disparate nature from genomics and medical information, to life-style and socio-economics. Normalisation and standards (see efforts by the European Open Science Cloud, ELIXIR, the Global Alliance for Genomics and Health and others) are essential to be able to answer complex medical questions. Still, large part of the information required to answer medical questions will be unstructured, like the images and reports of the example (including multilingual issues). These problems can now be **addressed with a new generation of AI/ML methods**, i.e. Deep Neural Networks, trained with large data sets and exceeding computing power.

Finally, not all medical problems will be addressable by direct operations in databases. In this case, **simulations** from atomistic, cellular, organ to organism will be the best instrument to explore complex situations. The **Medical Digital Twin**, even if not currently reachable, remains as a powerful paradigm of future medical support systems which arrival is worth accelerating.

4) Develop data analytical capabilities, including AI/ML, NLP and simulations, applied but not limited to, images, devices, textual sources (i.e. Natural Language Processing on EHR and other text medical sources, addressing multilingual issues) as part of decision support systems.

5) Explore the AI/ML technologies for training with distributed confidential data sources (federated learning)

c) Many medical questions, like image comparison or simulations, will require intense computation, while others can be solve with local devices. The development of technical solutions connecting data and software with the most adequate computational environment, from HPC to edge devices, in a transparent mode to the users will require the development of a **computing continuum ecosystem.**

6) Develop and integrate technologies to make possible the computation at the different scales from edge to HPC, fostering the collaborations between providers and IT developers including IT and software companies.

d) The support systems helping in diagnosis, like the one in the example providing matching imaging for a given critical case, will have to provide **reproducible and verifiable answers,** as well as being able to report what is the evidence used to support the proposed decisions (e.g., what are the key aspects by which the images in the example are considered similar enough to base a diagnosis (what is called **Explainable AI)**. Therefore, the **technology to benchmark and certificate** has to be developed in parallel with the medical support systems**.**

7) Develop systems for the simulation of biomedical systems at different levels from atomistic, cellular, organ to organism and for integrate of different level (multilevel simulations).

---

[5] For example the **European AAI authentication system** (https://elixir-europe.org/services/compute/aai)
[6] Publications in **Homomorphic Encryption** see for example: https://science.sciencemag.org/content/362/6412/347

8) Implement validation and certification strategies following the regulations and needs of different therapeutic areas[7], with particular attention to the development of explainable AI systems.

e) The new generation of decision support systems has the potential to **impact directly the economy of the medical systems, providing a more homogenous, reliable and egalitarian diseases diagnostic and decision support systems**.

9) IMI projects to address the social, regulatory and commercialization issues of the health data and data-related technologies.

10) Build capacity and provide training for the use of the new data technologies in collaboration with other European training efforts, professional and patient associations.

On behalf of the Scientific Committee

Isabelle Bekeredjian-Ding, Chair

[7] The ELIXIR OpenEbench (https://openebench.bsc.es/dashboard) provides an example of an ongoing effort to provide an open, live benchmark environment in the area of bioinformatics.