EU-STANDS4PM
standards for in silico models for personalised medicine

A European standardization framework for data integration and data-driven *in silico* models for personalized medicine – EU-STANDS4PM

Action plan on the concerted use of domain-specific standards for complex workflows in personalised medicine and systems medicine

Deliverable 1.3

Imprint

Authors

Arshiya Merchant and Niklas Blomberg
ELIXIR, EMBL-EBI

Publisher

EU-STANDS4PM administrative office on behalf of the EU-STANDS4PM consortium,
Forschungszentrum Jülich GmbH, Project Management Jülich, Germany

Contact: Marc Kirschner (m.kirschner@fz-juelich.de)

TABLE OF CONTENTS

OUTLINE

This report is based on an assessment on current personalised medicine and systems medicine projects and their use of data, standards and guidelines, as well as for their needs and gaps in standardisation. It puts a specific focus on analysing the interoperability and scalability of data and metadata standards specific for field domains (e.g. clinical data, systems medicine, bio banking, etc.) or applied technology domains (e.g. genomics, proteomics, metabolomics or modelling, etc.) to ensure cross-domain and cross-technology data integration. The report will provide recommendations for the concerted use of domain-specific standards for complex workflows in personalised medicine and systems medicine.

Introduction

What is a personalised medicine workflow?

A personalised medicine approach to improve patient health typically starts with identifying which aspect of health is to be addressed and modelled, for example predicting onset of disease, severity of a particular disease, or response to treatment(s) (1). Substantial efforts and funding, both within the private and public research sector, have been put towards modelling health and disease (2).

The next step typically is identifying relevant data and their respective sources. These different types of data from different sources often have to be harmonised, a task that is made easier when common data standards have been used. Once this has been completed, modelling to predict the clinically relevant questions will take place. In order to be used in a clinical setting, models then have to be validated in an independent setting, to ensure patient safety. Once that has been completed, the models can be used in a clinical environment to help improve patient health (1).

Challenges utilising a complex workflows for personalised medicine

Utilising workflows for personalised medicine becomes difficult, as accessing sensitive human data (from respective sources), especially across jurisdictional boundaries, can be quite challenging. These challenges are further amplified when conducted at a large scale, across multiple nations (2). Furthermore, the harmonisation of data is dependent on the use of appropriate standards (3). These standards ensure interoperability of data and metadata, higher quality of data, and are key drivers in the reproducibility of results. The appropriate integration of data and subsequent reproducibility of results increases overall trust in the models and ensures patient safety when translated into clinical settings (1). Thus, there is a need for broadly applicable standards for both the data, as well as the tools used *in silico* modelling for personalised medicine, that are compliant with national and international legal and ethical regulations that allow interpretation of a variety of health data through *in silico* methodologies to advance personalised medicine (1).

Aims of the report

The aim of this report is to evaluate the role of standards within a European federated ecosystem that enables the re-use of health data across borders to ultimately support the use of modelling in the context of personalised medicine.

It will briefly outline the current state of standards in the context of personalised medicine by evaluating the current use of data, standards, and guidelines in ongoing or recently finished personalised medicine and systems medicine projects. It will outline the current gaps and needs and provide corresponding recommendations for the research community and funding bodies to promote the concerted use of domain-specific standards for complex workflows in personalised medicine and systems medicine.

It is important to note that this report will be discussing the role of standards for the effective re-use of health data for secondary purposes, such as research and policymaking purposes. The use of health data for primary purposes such as the use of routinely collected health data in the delivery of healthcare is considered out of remit for this report and is covered by the European Health Data Space 1.

Background

The EU-STANDS4PM consortium, along with other European Commission (EC) funded projects, have conducted a considerable body of research on the type and use of standards within health research. Table 1 explores a select few outcomes from them.

Table 1: Examples of standards within health research.

Project	Output	Relevance
EU-STANDS4PM	EU wide mapping effort, focus on international databases, collections, & registries.	This report provides an overview of which data types are relevant for different data types pertinent to personalised medicine research, as well as highlights the importance of standards to govern these databases, collections, and registries.
EU-STANDS4PM	EU wide mapping report on good practice examples for integrating phenotype and large scale data.	This report provides national and EU-case studies (good practice examples) for integrating patient derived data, such as phenotype and large scale data, for *in silico* modelling in personalised medicine. This report provides a list of common standards currently relevant to personalised medicine (Table 1.0).
EU-STANDS4PM	Towards standardised guidelines for *in silico* approaches	This report provides a summary of the COMBINE workshop which addresses crucial requirements with respect to the development of data and model standards as well as data integration tasks in research and clinic, including ethical and legal aspects. Additionally, Table 2, within this article provides a list of common standards, including omics, imaging, and tools and analysis pipelines relevant to personalised medicine and *in silico* approaches.
Beyond 1 Million Genomes - Work Package 3	Documented best practices in sharing and linking genetic and phenotypic data.	An iterative document that documents the current best practices to integrate genomic and phenotypic data.
Beyond 1 Million Genomes - Work Package 4	Federated Data Access Rare Disease - Proof of Concept (https://youtu.be/6MtIJA4xXdU)	A demonstrator that outlines how to bring together the GA4GH standards and infrastructure services and components to enable federated access to genomic data.
1+ Million genomes	Trust Framework *	A framework with five main components that enable genomic data to be shared across borders and implemented into healthcare systems.
European Health Data Space Pilot (2)	Metadata catalogue*	Metadata catalogue for five use cases with varying types of data types.

*to be published

The above outputs from various EC funded projects on the role of standards in enabling federated access to data for personalised medicine research outline the ongoing efforts, including providing expansive lists of domain-specific standards relevant to personalised medicine and systems medicine, as well as

the gaps in the concerted use of them. The gaps and corresponding recommendations are explored in further detail below.

CURRENT GAPS & CORRESPONDING RECOMMENDATIONS

Formal, Commercial, and Community Standards

Standardisation plays an important role in research and innovation (R&I) investment agendas as it helps pave the way for large-scale deployment of new and strategic technologies. Standards help to validate and spread the scientific discoveries and inventions towards healthcare and furthering personalised medicine (4).

Formal standards are created by official international standardisation bodies he International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), or the International Telecommunication Union (ITU) and take several years to develop. A formal standard, once completed and released, is internationally respected and recognized as state of the art – also from a legal perspective (1).

Creating formal standards has been a central objective of the EU-STANDS4PM consortium as demonstrated by initiating the ISO Technical Specification “Biotechnology — Recommendations and requirements for predictive computational models in personalized medicine research — Part 1: Guidelines for constructing, verifying and validating models" (ISO/DTS 9491-1; 5). “This document defines challenges and requirements for predictive computational models constructed for research purposes in personalized medicine. It specifies recommendations and requirements for the setup, formatting, validation, simulation, storing and sharing of such models, as well as their application in clinical trials and other research areas.”

Oftentimes companies and organisations themselves drive the process of creating formal standards and these could be referred to as commercial standards. These vary widely and depend on who created them (e.g. a small research group or a large company) and what the intention was when they were created (e.g. to provide interoperable data for customers or to secure their own market advantages). They are also taken into account during the creation of formal standards.

However, due to the processes required to create formal standards, their creation takes a considerable amount of time, approximately 3-5 years, and requires a substantial amount of resources, including the experts’ time (1, 6). Additionally, the long development time also runs the risk of the standard being obsolete by the time it is officially approved.

Community standards on the other hand, usually reflect the results of a specific user group and are created by individual enterprises or communities. Community standards vary across various topics and there is no prescribed process for creating, agreeing, and consensus-building – but also no time frame. Therefore, community standards are usually available within a relatively short time and can be adjusted to the user (1).

Due to the short amount of time needed to develop community-driven standards, as well as the flexibility for adjustments and enhancement, often results in several versions of the same standard (8). Additionally, community-driven standards, whilst representing a portion of the scientific community do not represent the entire community and it is important to note that even community-based standards need to be disseminated throughout and adopted by the community (1).

Lastly, there is a competition between standard development organisations (SDOs) which results from the natural evolution and expanding scope of the work. This competition forces implementers to choose among multiple options and requires an additional step of mapping between standards using an interface engine for interoperability when combining data from different sources.

Recommendations

⇨ **Balancing timing:** As mentioned above the creation of formal standards takes a considerable amount of time, expertise, and funding. The long development process could potentially render the standard obsolete by the time it is released. Standards setting organisations need to create a balance between ensuring that the standards are mature and have been thoroughly investigated versus approving standards prematurely that may not cover all aspects. This is especially key in the context of personalised medicine.

⇨ **Proactive planning and resource allocation:** The creation of standards, especially formal standards, required significant investment from experts, in both time and resources. Funding bodies play a vital role in the standardisation landscape and it is recommended for them to ensure that adequate funding is budgeted for within research consortiums, or alternative forms of contributions to research groups to allow them to continue their efforts within this space.

To aid in the allocation of resources, funding bodies in collaboration with standards setting organisations and researchers should proactively plan for the management of health information, as currently the development of standards is done on a basis of need.

Standards – Data

Data collection and pre-processing: It is well established that biomolecular profiles can change markedly during sample collection, storage and pre-processing, limiting reproducibility and comparability of studies and collection. Variability may also make outcomes and diagnostics unreliable as the results from studies are not translatable to a routine clinical setting. European projects, such as Spidia4p (https://www.spidia.eu/), have developed procedures and standards to accurately capture such aspects, the widespread adoption of these standards and recommendations across European research performing organisations would be important for reproducible research and for generating large, comparable and high quality datasets that would be use.

Recommendations

⇨ **Research performing organisations and funders** should recognise the importance of standards in the quality and reproducibility of research results and consider how these aspects can be funded and incorporated into future research programs.

Data Visibility & Reusability

Interoperability in the context of personalised medicine and systems medicine research can be broadly categorised by interoperability of the information (data), interoperability of the tools utilised, interoperability of the models used, and the clinical guidelines and healthcare business processes that enable the re-use of the healthcare data (9).

Focusing on the interoperability of data, health data for reuse, has been difficult to find. A key concern of researchers is the availability of data. (10). Following that, another key challenge has been that when the data becomes available there is a distinct lack of accessibility as a result of ELSI issues (further explored below) owing to the sensitive nature of the data in question, and a lack of transparency of usage which impacts the reproducibility of the results (1,6).

These issues are more thoroughly explored in the EU-STANDS 4PM consortium deliverables[1]:

- D1.1: EU-wide mapping report with focus on international databases, collections and registries (8)
- D1.2: EU-wide mapping report on good practice examples for integrating phenotype and large scale data (12)
- D3.2: Harmonization and integration of big data of relevance for personalized medicine into in silico modelling? – Recommendations for technically feasible, and ethico-legal sustainable avenues (13)

[1] EU-STANDS4PM deliverables are available for download under: https://www.eu-stands4pm.eu/publications

FAIR in the context of personalised medicine workflows

Application of the FAIR standards is critical for personalised medicine workflows, starting from the infrastructure that would support these processes and the data that will be utilised to eventually improve care.

The overall goal is to provide secure, standardised, documented and interoperable services under a common framework in which data sets remain within appropriate jurisdictional boundaries; whereas, metadata (for example, data set descriptions) are centralised and searchable through a common application programming interface (API). After data discovery, access to the data themselves can be requested from the source, for example, by applying to a data access committee, to establish agreements for data use.

As demonstrated in previous deliverables (Table 1), data and metadata collected by disparate cohorts varies greatly, and the information is collected for different purposes. Standardisation and interoperability of these data is critical to prevent seclusion of data in silos and can be achieved through the application of FAIR principles. Projects that aim to build EU wide research infrastructures should carefully consider what is the appropriate level of standardisation - will community standards suffice or there is a need for more formal standards?

Data Integration

Health data is an umbrella term for many different varieties of data types. *In silico* modelling in the context of personalised medicine research requires the input of a variety of different data types (depending on the type of research being carried out).

The issue of integration of different data types is more thoroughly discussed in Brunak et al. (1) and D1.2 as they outline the different methods of integrating data, their challenges including establishing semantically consistent disease annotations and medical vocabulary, handling different types of patient populations and overcoming highly diverse registration procedures of measurements and interventions, followed-up with use cases supporting the different methods.

Recommendations

⇨ **Availability of data sets:** Comprehensive and complete dataset on a patient in specific areas of interest should be available, further highlighting the importance of integration of data (1, 15). Precision medicine research would greatly benefit from a combination of basic, translational and clinical research. These methods are explored further in D1.2 (8) including individual level integration, integration of variables, and integration of unstructured data.

⇨ **Data integration:** Strategies for the integration of heterogeneous and typically disconnected and unstructured data sources (15) should be investigated. Being able to combine data from different sources necessitates a common vocabulary, further encouraging interoperability (16). The processes governing the integration would be made easier by adhering to standards developed either formally or community-driven (17).

Standard - Tools

The increasing need for storing/analysing/visualising health data led to the evolution of software tools. We have moved from simple scripts, to sophisticated software, to machine learning, to databases and their management system and ultimately to workflows that wrap and execute everything together. Tools range from databases, read-across and (Q)SARs, to predictive software and complex machine learning algorithms. Some methods are simple and intuitive, while others require much more expertise to develop and use. Recent efforts have been directed towards developing more transparent, mechanistically-driven models (11).

In the context of personalised medicine, there are rising concerns on multiple aspects regarding the development and execution of tools. These can only be tackled through a widely accepted standardisation effort.

Recommendations

Perhaps the biggest necessity in the tools development is that they have to be very clear on the underlying mechanism by which outputs are produced. Decisions enclosed in black-boxes – a trend very often seen in neural networks - are unethical and cannot be sustained. This also entails considering insightful visualisations in the analysis process (18).

⇨ **In silico work flows:** In silico-modelling is a very niche domain, completely disconnected from people outside the computational biology field. This means that the workflow should be open to understanding it as well by having access to all related metadata (17, 18). For example, it is important that training materials are linked to the respective tools of the workflow. These training materials can be considered metadata. The more rich the tools/workflow metadata, the more complete and transparent and open research becomes. By using common ontologies in the tools/workflow development and description process, it becomes easier to aggregate and interconnect all peripheral information for that workflow by using data mining for example. Peripheral information could be training relative to the respective analysis, e.g. how mass-spectrometry that produces an output is used as a feature to the model. Development of tools should follow the FAIR principles to ensure reproducibility. Workflows should be runnable and provide reproducible information, irrespective of the platform that is being used (18). There is a great effort in containerisation of software for this issue.

⇨ **Future considerations**: developers should follow FAIR guidelines for software principles and the community of tool developers should enable continuous benchmarking processes and metadata coordination of these complex workflows.

Standards - Models

Standards are essential to enable FAIR digital personalised medicine. They enable all Findability, Accessibility, Interoperability and Reusability (FAIR) of data. They also increase the quality and value of data by representing their meaning in data models unambiguously and understandable for both humans and machines. This is extremely important for integrating data across different scales and resolutions necessary for systems medicine studies and for data-driven computational methods such as Machine Learning (ML) workflows used for *in silico* modelling and prediction of new hypotheses. However, the complexity of health data, and heterogeneity of health data models and domain-specific standards poses a serious barrier to accelerate innovations in personalised medicine.

Recommendations

⇨ **FAIR guiding principles:** FAIRification is the process to make data FAIR by adopting FAIR implementations and community standards. To lower the barrier of *in silico* prediction on health data in the context of personalised medicine, FAIRification should be performed with a focus on biomedical and ML standards, i.e. domain-specific FAIRification. However, it is not trivial how to reach community consensus on what FAIR implementations and standards to adopt. We recommend establishing as soon as possible an objective procedure to identify and agree upon community FAIR implementation choices and standards, and consolidate them as recommendations for the community. In addition, developing domain-specific FAIR Maturity Indicators such as those developed within the Rare Disease community, help on both FAIR evaluation of data resources and guidance on what standards and FAIR implementations need to be implemented to be FAIR within the health community for biomedical analysis.

Legal, Regulatory, and Ethical Considerations

Federating large volumes of sensitive health data across internationally distributed virtual computing environments presents formidable challenges in assuring data integrity, service availability, and individual privacy (11). These challenges are further explored in the D3.2, as well as work published by Work Package 2 of the Beyond 1 Million Genomes consortium, and GA4GH.

In order to share sensitive health data across jurisdictional boundaries across the EU, one has to navigate GDPR, as well as individual country regulations and policies governing data sharing.

In addition, there is a severe lack of understanding of the terminology associated with data sharing (also further explored in D3.2 as well as the work package 2 scoping document by the Beyond 1 Million Genomes consortium)

In silico modelling also raises concerns regarding the type of information that must be supplied to research subjects to meet GDPR consent requirements. This can be a challenge for *in silico* models where the research hypothesis is unclear (1).

Recommendations

⇨ **EU-wide legal and technical interoperability**: The European Health Data Space proposed legislation (as of 2 December 2022, the Proposal for a Regulation on the EHDS is being discussed within the Council of the European Union) and accompanying federated infrastructure should provide a legal and technical framework for researchers to share and access health data for secondary purposes. However, care must be taken to ensure that the legislation and infrastructure are implemented in a manner that promotes legal and technical interoperability between nations.

Additionally, D3.2 delves deeper into (twelve) recommendations that outline solutions to the challenges arising from data integration for *in silico* modelling for personalised medicine. Some of these twelve solutions can be established within the existing national and European legal and ethical framework. Others will need adjustments or amendments of national or EU-law.

CONCLUSION

Standardisation efforts should also be fully fundable to ensure that appropriate and sufficient resources are made available to the scientific communities for developing standards that the researchers then could apply consistently to their workflows. This ensures establishment of standards that reflect best practice in their domain. Data processing, documentation, and subsequent sharing thereby become integral, obligatory deliverables of funded projects, included in the budget and planning. Data sharing and documentation thereby become less onerous than currently, where they are un-funded and altruistic.

Additionally, policy makers and funders must continue to allocate financial resources to programs that support the development of new research software and the maintenance of research software that has a large user base and/or an important role in a research area. By providing the resources that are necessary to adhere to best software development practices, policy makers and funders can increase overall software quality and usefulness.

Policy makers and funders should provide programs and funding opportunities that encourage both researchers and research support professionals (such as Research Software Engineers and Data Stewards) to utilise best practices to develop better software faster**.** In order to make research software understandable and reusable, it must be produced and maintained using standard practices that follow standard concepts, which can be applied to software ranging from researchers writing small scripts and models, to teams developing large, widely-used platforms.

REFERENCES

1. Brunak, S., Bjerre Collin, C., Eva Ó Cathaoir, K., Golebiewski, M., Kirschner, M., Kockum, I., Moser, H. & Waltemath, D. (2020). Towards standardization guidelines for *in silico* approaches in personalized medicine. *Journal of Integrative Bioinformatics*, *17*(2-3), 20200006. https://doi.org/10.1515/jib-2020-0006
2. Andreas Bender, Angela Poschlad, Stefan Bozic, Ivan Kondov, A Service-oriented Framework for Integration of Domain-specific Data Models in Scientific Workflows, Procedia Computer Science Volume 18, 2013, Pages 1087-1096,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2013.05.274.
3. Ola Spjuth, Andreas Karlsson, Mark Clements, Keith Humphreys, Emma Ivansson, Jim Dowling, Martin Eklund, Alexandra Jauhiainen, Kamila Czene, Henrik Grönberg, Pär Sparén, Fredrik Wiklund, Abbas Cheddad, þorgerður Pálsdóttir, Mattias Rantalainen, Linda Abrahamsson, Erwin Laure, Jan-Eric Litton, Juni Palmgren, E-Science technologies in a workflow for personalized medicine using cancer screening as a case study, *Journal of the American Medical Informatics Association*, Volume 24, Issue 5, September 2017, Pages 950–957, https://doi.org/10.1093/jamia/ocx038
4. *Standards Drive Innovation*. European Commission. (2021, August 4). Retrieved September 29, 2022, from https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/standards-drive-innovation-2021-08-04_en
5. *ISO/DTS 9491-1: Biotechnology — Recommendations and requirements for predictive computational models in personalized medicine research — Part 1: Guidelines for constructing, verifying and validating models*. https://www.iso.org/standard/83516.html
6. Collin, C. B., Gebhardt, T., Golebiewski, M., Karaderi, T., Hillemanns, M., Khan, F. M., Salehzadeh-Yazdi, A., Kirschner, M., Krobitsch, S., & Kuepfer, L. (2022). Computational models for clinical applications in personalized medicine—guidelines and recommendations for data integration and model validation. *Journal of Personalized Medicine*, *12*(2), 166. https://doi.org/10.3390/jpm12020166
7. Myers CJ, Bader G, Gleeson P, Golebiewski M, Hucka M, Nove`re NL, et al., editors. A brief history of COMBINE. 2017 Winter Simulation Conference (WSC); 2017 3-6 Dec.2017.
8. EU-STANDS 4PM report: *EU-wide mapping report on good practice examples for integrating phenotype and large scale data (D1.1).* https://www.eu-stands4pm.eu/publications
9. K. Heubusch, "Interoperability: What it Means, Why it Matters," Journal of AHIMA, vol.77, n°. 1, pp. 26-30, January, 2006.
10. Leipzig. (n.d.). *RDA COVID-19 recommendations and guidelines on Data Sharing · issue #18 · Leipzig/metadata-in-RCR*. GitHub. Retrieved December 5, 2022, from https://github.com/leipzig/metadata-in-rcr/issues/18
11. Madden JC, Enoch SJ, Paini A, Cronin MTD. A Review of In Silico Tools as Alternatives to Animal Testing: Principles, Resources and Applications. *Alternatives to Laboratory Animals*. 2020;48(4):146-172. doi:10.1177/0261192920965977
12. EU-STANDS 4PM report: EU-wide mapping report with focus on international databases collections and registries (D1.1). https://www.eu-stands4pm.eu/publications
13. EU-STANDS 4PM report: Harmonization and integration of big data of relevance for personalized medicine into in silico modelling? – Recommendations for technically feasible, and ethico-legal sustainable avenues (D3.2). https://www.eu-stands4pm.eu/publications
14. Boehm, K.M., Khosravi, P., Vanguri, R. et al. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer 22, 114–126 (2022). https://doi.org/10.1038/s41568-021-00408-3
15. Kraus, J.M., Lausser, L., Kuhn, P. et al. Big data and precision medicine: challenges and strategies with healthcare data. Int J Data Sci Anal 6, 241–249 (2018). https://doi.org/10.1007/s41060-018-0095-0
16. Bhartiya, S., & Mehrotra, D. (2014). Challenges and recommendations to healthcare data exchange in an interoperable environment. *Electronic Journal of Health Informatics*, *8*(2), 16.
17. Carole Goble, Robert Stevens, State of the nation in data integration for bioinformatics, Journal of Biomedical Informatics, Volume 41, Issue 5, 2008, Pages 687-693, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2008.01.008
18. Barker, M., Chue Hong, N.P., Katz, D.S. et al. Introducing the FAIR Principles for research software. Sci Data 9, 622 (2022). https://doi.org/10.1038/s41597-022-01710-x

ACKNOWLEDGEMENTS

EU-STANDS4PM is funded by the European Union Horizon2020 framework programme of the European Commission Directorate-General for Research and Innovation under Grant Agreement # 825843.

The current report is part of WP1 Data sources and standards for predictions in personalized medicine, and Deliverable 1.3 *Action plan on the concerted use of domain-specific standards for complex workflows in personalised medicine and systems medicine.*

CONTACT

Arshiya Merchant
EMBL-ELIXIR
E-mail: arshiya.merchant@elixir-europe.org

Marc Kirschner (Coordination)
Forschungszentrum Jülich GmbH
Project Management Jülich (PtJ)
52425 Jülich, Germany
Phone: +49 2461 61-6863
E-mail: m.kirschner@fz-juelich.de
www.eu-stands4pm.eu

Sylvia Krobitsch
Forschungszentrum Jülich GmbH
Project Management Jülich (PtJ)
52425 Jülich, Germany
Phone: + 49 30 20199-3403
E-mail: s.krobitsch@fz-juelich.de


EU-STANDS4PM
standards for in silico models
for personalised medicine