Data harmony and standards: data must be processed, described and stored by uniform means

A lot of genetic data and clinical data is easily generated from measurement instruments.

It is important, however, to decide in good time how and in which format the raw data is stored and how the postprocessed data is classified and described, the measurement event included.

It is important to determine metadata, i.e. the information that describes the data, by the exact same means in all research institutions and laboratories around the world. Otherwise, we cannot obtain the maximum benefit of the data for research, as it cannot be linked to data produced elsewhere.

"Even within our own research group it can become chaotic if, for example, the same files have not been used, meaning that the data is not comparable", says professor **Aarno Palotie** at the Institute for Molecular Medicine Finland (FIMM), University of Helsinki.

Owing to international cooperation, some standards have already been created. The VCF (Variant Call Format) format determines the text file used in bioinformatics when storing genetic sequence variation data. BAM (Binary Alignment/Map) is a format that can be converted to a readable text format.

"it allows more than 1,000 organisations to work together in creating common principles for data processing and distribution."

Founded in 2013, GA4GH (Global Alliance for Genomics and Health) is an international alliance of more than 500 organisations representing the bioinformatics, health care and IT industries with the aim to create standards for data that is distributed for research purposes. In November 2017, ELIXIR and GA4GH decided to launch cooperation. The agreement gives the ELIXIR infrastructure an opportunity to influence the creation of international standards. The agreement is related to the GA4GH Connect project, the purpose of which is to introduce data standards in clinical patient work by 2022.

Aarno Palotie finds cooperation between ELIXIR centres and GA4GH important because, in addition to standards, it allows more than 1,000 organisations to work together in creating common principles for data processing and distribution.

"The ELIXIR centres have extensive networks in their respective countries and can influence the local practices."

Legislators need to understand the various needs for using data

There is still work to do to achieve uniform data processing, analysis and principles of use. In Finland, the aim is to enable exploitation of genome data in patient health care. The purpose is to create a national genome data resource which will be maintained by Genome Centre Finland.

"In Finland, the aim is to create progressive laws for the linking and utilisation of clinical and genome data of the population."

According to Palotie, the various uses of data must, however, be considered in the linking of clinical data and genome data used in research.

"No errors are allowed in patient-oriented analysis. Messing up with samples cannot be tolerated. The data must be available in identical form and easily accessible if it is used for clinical decision-making. Scientific genome data must, in turn, be flexible, quickly accessible and available in various formats. Research will advance only by flexible means."

The researchers involved in the FinnGen project have had to deal with many kinds of agreements. The data protection regulations require extremely stringent protocols agreed upon in advance which, according to Palotie, is in contradiction with the research ideology.

"Basic research just fails to progress if it follows protocols determined from the outside. In research, the processes are modified and applied according to how the data is produced. The aims are different than when utilising clinical genome data."

> **"The aims are different than when utilising clinical genome data."**

As the requirements vary, Palotie says that parallel routes are needed for the exploitation of these two different types of data. Legislation should further specify how genome data created for clinical purposes can also be utilised in research.

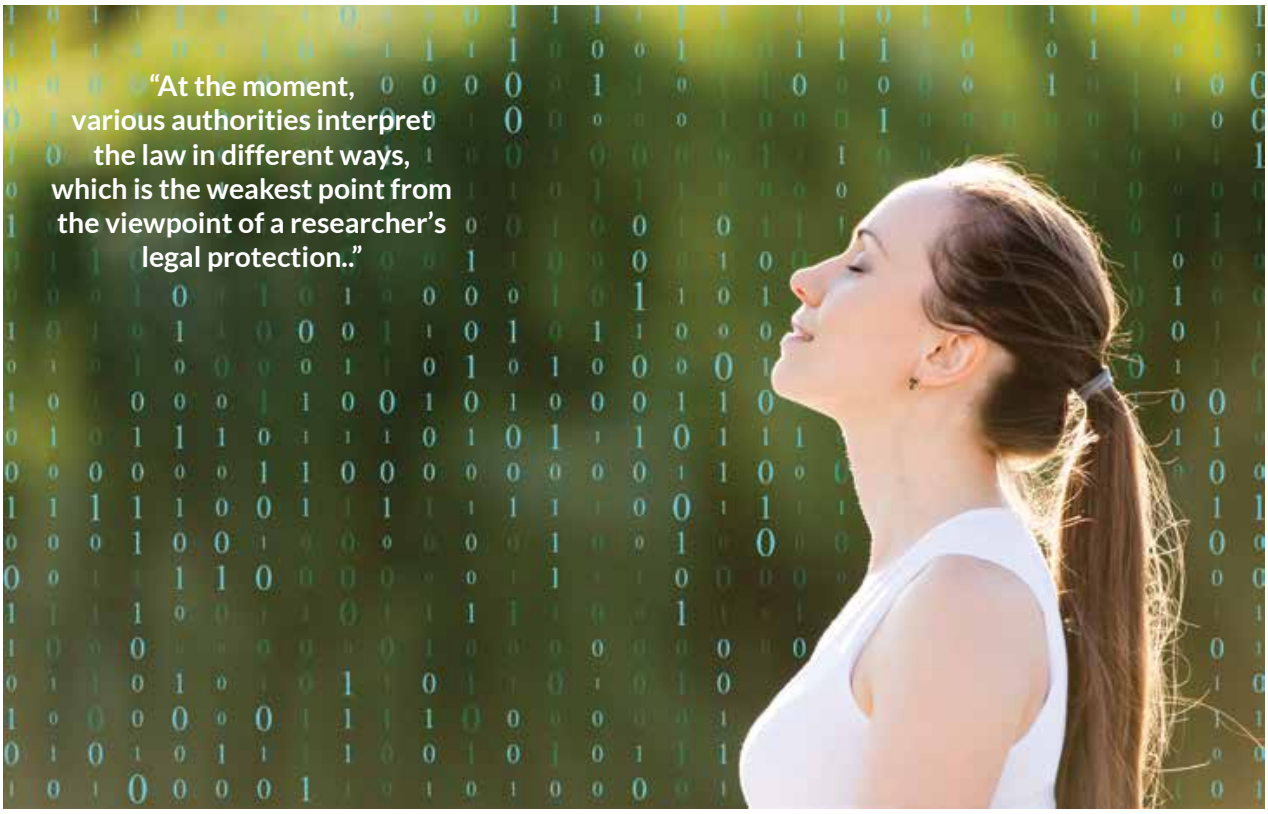
"A consensus should be reached also on how the data generated in research could, in some situations, be sensibly used for clinical decision-making. The situation is unclear at the moment in the Finnish Biobank Act."

Naturally, it is important to take care of the necessary data protection, but excessive or tangled regulation of data use is troublesome from research point of view. The European regulatory environment is partially broken. Palotie mentions the mining of metadata as an example.

"A researcher wants to know, for example, how many individuals there are in the Finnish biobanks with a certain genotype, illness and age. Ideally, we would have a portal, which would give this piece of information in real time. A researcher cannot access the individual data which the computer is processing in the depths of the system. When utilising personal data, which is subject to stringent protocols in the EU, the starting point is that the data must not be processed. Research makes an exception here, but when strictly interpreted, it may even require a separate permit process each time."

Utilisation of data is challenging because the regulatory environment has mainly been interpreted from the viewpoint of data protection, and not from that of the health benefit to an individual.

"The regulatory environment should be developed in such a way that all authorities interpret the legislation in the same way, which would also allow using the portal I have proposed. The use of this kind of portal, if constructed in an appropriate manner, does not pose a threat to data protection by any means whatsoever", says Palotie and points out that the metadata regulations



are looser, for example, in the United States than in Europe.

“The European permit processes have been characterised as a bureaucratic farce. It may take years to receive a permit”, Palotie sighs.

Palotie wishes that the new legislation created in Finland that relates to using genome data and secondary use of register data would accelerate the permit processes and clarify the regulations concerning the use of data.

“The data policy needs to be clear and the data available to everyone. At the moment, various authorities interpret the law in different ways, which is the weakest point from the viewpoint of a researcher’s legal protection. Hopefully the new laws will remedy this situation. I also hope that the updating of the Finnish Biobank Act due to the General Data Protection Regulation of the EU (GDPR) is aligned with other new legislation.”

Ari Turunen

MORE INFORMATION:

Institute for Molecular Medicine Finland (FIMM)
The mission of the Institute is to advance new fundamental understanding of the molecular, cellular and etiological basis of human diseases. This understanding will lead to improved means of diagnostics and the treatment and prevention of common health problems. Finnish clinical and epidemiological study materials will be used in the research.

CSC – IT Center for Science
is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.
http://www.elixir-finland.org
http://www.elixir-europe.org

ELIXIR FINLAND
Tel. +358 9 457 2821s ● e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR HEAD OFFICE
EMBL-European Bioinformatics Institute
www.elixir-europe.org