Disease prediction models are becoming more accurate thanks to computational methods

Computational methods can now be used to deduce from data sets as to who is at risk of developing, for example, diabetes or cancer. Laura Elo and her research group develop methods which are used to find different predictive markers for diseases. Combining clinical data with molecular data can also provide valuable information about suitable drug treatment.

Research conducted on human biology produces a lot of new data for researchers to study. DNA sequencing generates an individual's genetic profile. RNA sequencing, in turn, provides measurement data on the activity of genes. It tells which genes are expressed and possibly produce proteins in the cells at any given time.

Thousands of different molecules and their interactions can be measured from a tissue sample. For example, it is possible to study different active forms, or transcripts, of a gene. When the goal is to determine the function of proteins or their deviations in connection with diseases, it is called proteomics. Mass spectrometers are used as aids to measure molecular mass.

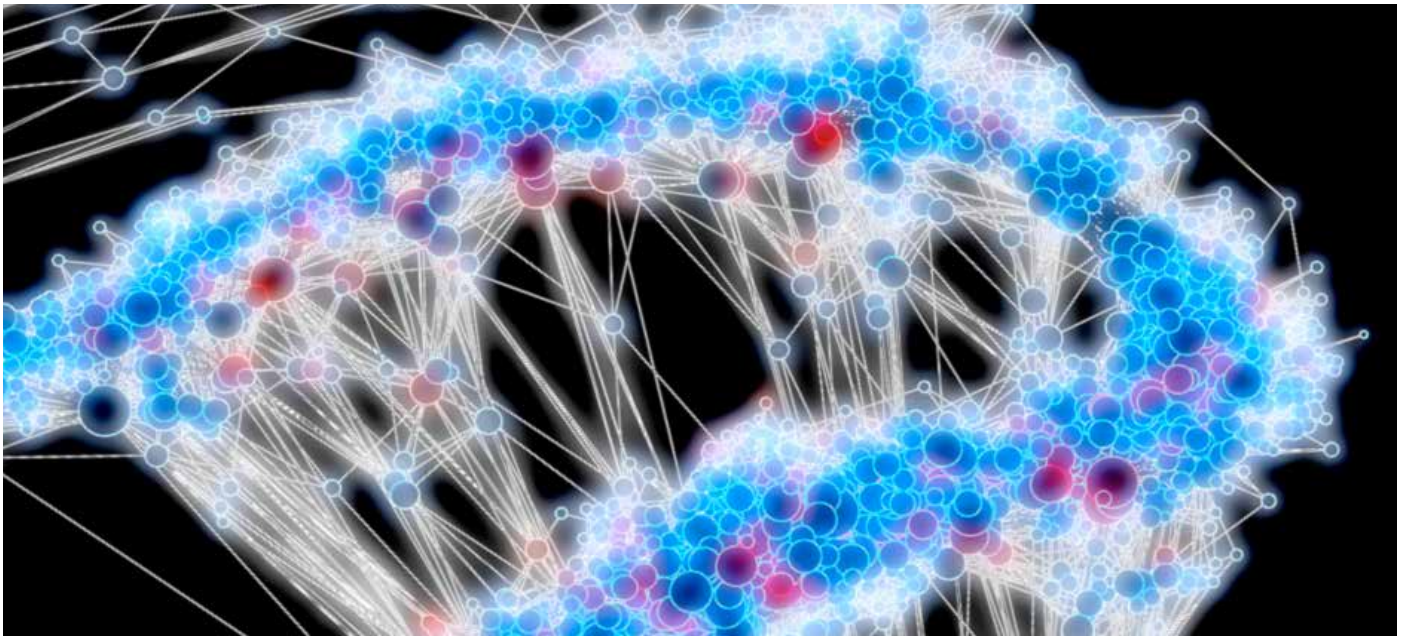
Laura Elo, Research Director of Bioinformatics at the Turku Centre for Biotechnology, and her research group develop modelling methods that allow the measurement data collected in follow-up studies to be utilised to determine disease risk on an individual basis.

"I started my career as a mathematician at a time when bioinformatics was still a marginal field. I became excited about biology and medicine back then", Elo says.

One important material for the researchers is the data collected from different populations. The studies use data stored in Auria Biobank in the Turku region as well as data obtained from elsewhere in Finland and from other countries. The electronic medical records also have a lot of data collected from patient care that can be used in research, subject to consent. However, the data alone is not enough to determine the emergence and development of diseases. Computational methods and models are required to make comprehensible interpretations from data masses. The aim is to develop functional models for use by doctors.

"Almost all of our research is related to medicine and the needs of doctors. One of our major goals is to provide practical tools for doctors. The data alone is not useful, unless it can be modelled and interpreted. In the future, our work will hopefully allow patients to be offered treatments that are increasingly individually targeted."

Effective treatment is always personal because drugs and treatment methods work in different ways for different individuals. A patient's treatment response is affected by a number of factors, the information about which is obtained, for example, through laboratory measurements. In addition to clinical variables related to the patient's health, there are many factors at the gene and protein level, which affect the efficacy of treatment methods. Mathematics helps with the analysis of data obtained about an individual.



Computational modelling has resulted in the creation of a network whose nodes represent individuals and the arches between the nodes the relationships between individuals.

"Biology is complex. One disease can actually present itself in many different ways at the molecular level, and different treatments can be effective for different people. A specific drug can cause serious adverse effects for some while being ineffective for others. Computational methods allow us to predict, who will suffer from the adverse effects and who are likely to benefit from the treatment. We mathematicians can help medical scientists to identify the key predictive factors", Elo says.

The model must also be suitable for new data

Development of mathematical models requires large volumes of data as their raw material. For example, some of the predictive models have been developed using clinical patient data from the US, but they are also suitable for the patient data of the Turku University Hospital.

"When a sufficiently large amount of genomic and clinical data is obtained, they can be combined and the modelling phase can start. This is only possible if the description of the data, metadata, is in order."

Many things have to be taken into account in the development of models. It is important to assess the prediction ability of the model in advance. Models easily become overfitted for the data that is used to create them. This means that the model is too well-suited for the data. Therefore, the predictive model works with one data, but the prediction is no longer good with new data. Validation is required to verify the model. This can be achieved, for example, by using patient cohorts from another hospital or country. Checking the model by using other patient data is important to allow for the general adoption of the model. Data from different biobanks helps with this.

"If the model is built and tested using the same data, you may get it to work almost perfectly in that data. However, it may not work on new individuals. Therefore, we strive to build models which predict the outcome as closely as possible but can still be generalised to new data."

The work of Laura Elo and her research group with modelling involves continuous experimentation and change.

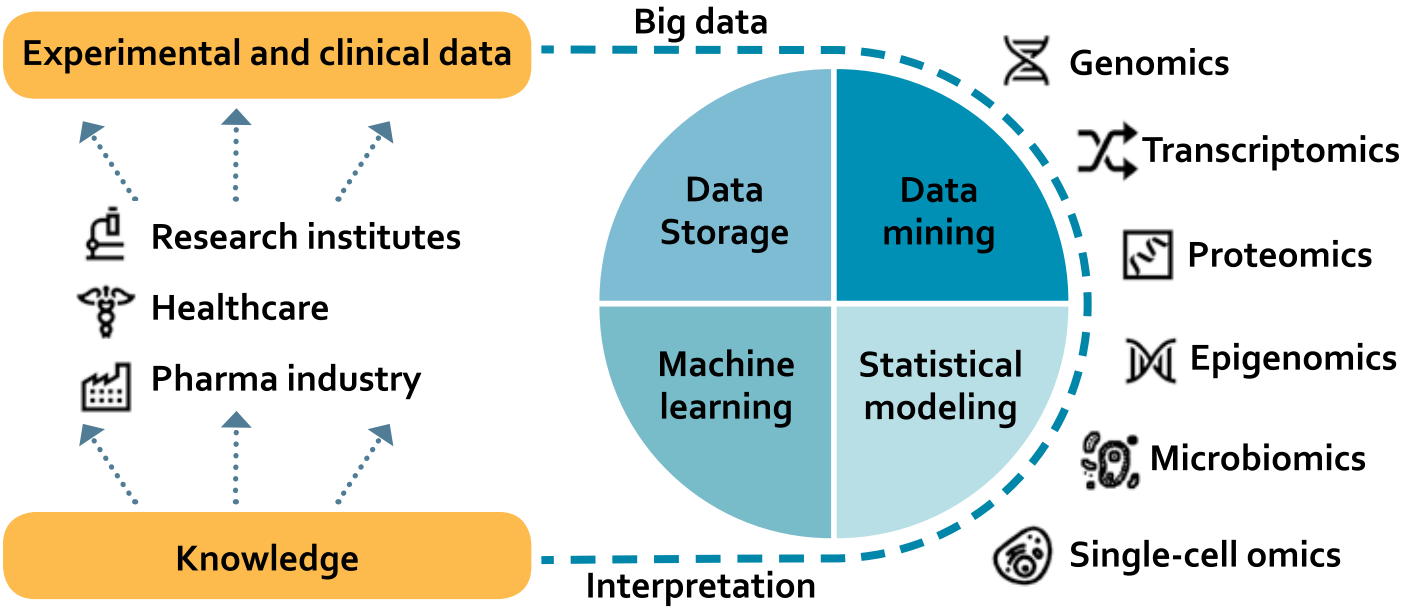
"After developing a model and showing that it works with certain data, the validation process is continued. We aim to find as many new data sets as possible to test the accuracy of the predictions produced by the model. You can always develop a model that works in one data. However, it is only after it has been verified in several data sets that the predictive model can be considered reliable enough to be given to doctors to support decision-making. The more widely the model can be tested, the better we can assess whether it only works for a specific population or if it is more universally applicable."

New factors are added to models and their effect on predictions is analysed. For example, linear, simplifying models are easy to understand and interpret in hospitals. However, sometimes the interactions between molecules are so complex that linear models do not work and, therefore, other solutions are needed.

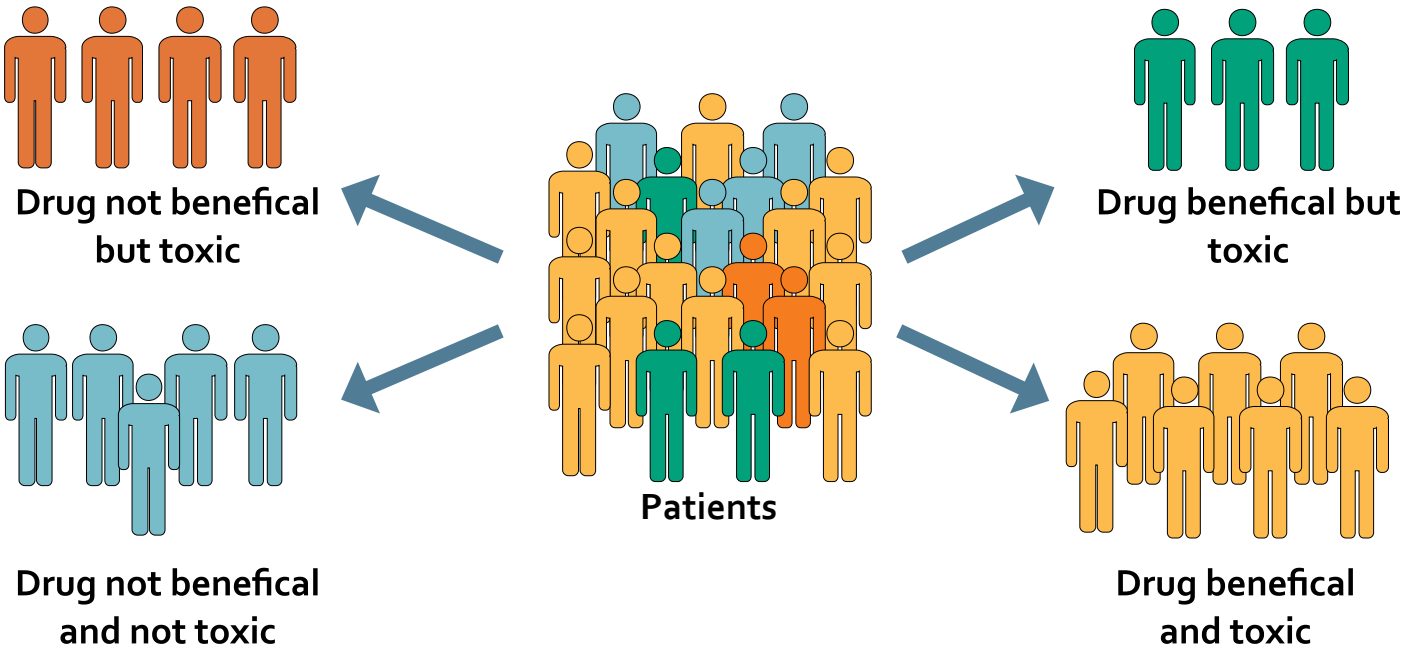
"The more new variables are added to a model, the more critical its validation becomes. An important question is understanding which variables are most significant for prediction and how their combinations provide the best predictions. You need to find balance for the model: it must be complex enough for prediction, but the model must not be overfitted to the data."

Predictive model for renal cell carcinoma

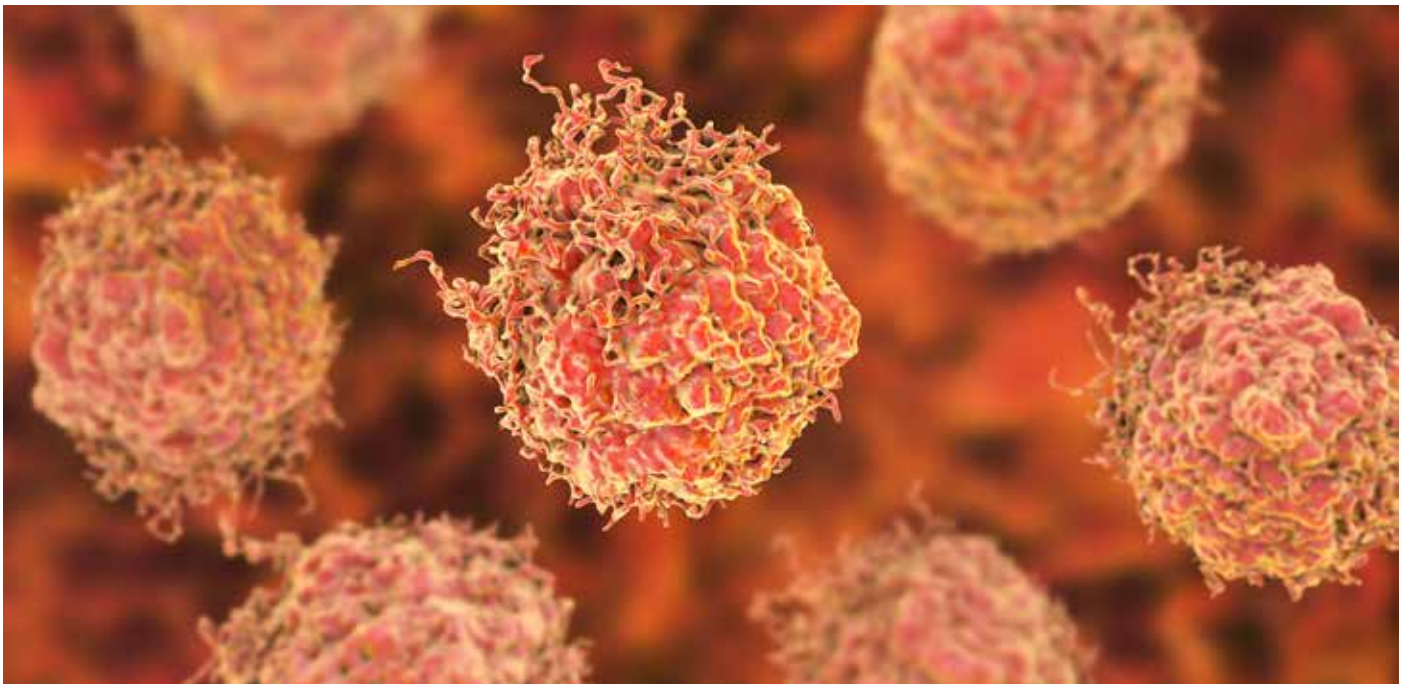
Laura Elo and her team have been involved in the development of predictive models for renal cell carcinoma. Renal cell carcinoma originates in the epithelial cells of the renal cortex. The prognosis of renal cell carcinoma is poor as 40% of patients die within five years.



The Medical Bioinformatics Centre develops computational data analysis tools and mathematical modelling methods for the needs of biomedical research. Special focus is put on the analysis and interpretation of the extensive measurement data produced by modern biotechnology (e.g. deep sequencing and mass spectrometry). The goal is to improve the diagnostics, prognoses and treatment of complex diseases, such as diabetes and cancer, in close cooperation with doctors and medical researchers.



The aim of personalised medicine is to identify factors that can be used to find the most suitable treatment strategy for each individual.



Prostate cancer that spreads metastases and is resistant to hormonal treatment is a malignant disease leading to the patient's death. The cytostatic drug docetaxel was introduced over a decade ago. However, approx. 10–20% of patients have side effects that force them to stop the treatment. International research groups created mathematical models that predict the side effects of cytostatic prostate cancer treatment for the Prostate Cancer DREAM 9.5 Challenge. The researchers developed a total of 61 models for the challenge, seven of which turned out to work and were awarded in the competition. A model developed by the joint research group of the University of Turku and the Turku University Hospital was one of the winning models. More information: Journal of Clinical Oncology Clinical Cancer Informatics: http://ascopubs.org/doi/abs/10.1200/CCI.17.00018

A new computational method can be used to find predictive markers from patient samples. The study found that the expression of 152 genes can predict the life expectancy of patients with renal cell carcinoma after surgery.

"The prognosis of renal cell carcinoma is usually good if the cancer is localized. On average, however, 50% of patients develop metastases after surgery. The goal is to predict as early as possible whether the patient's prognosis is good or bad in order to select the best treatment strategy."

Two different sets of data were utilised in the development of the predictive model. The gene expression data of more than 400 renal cell carcinoma patients were obtained from the international Cancer Genome Atlas (TCGA) database. The model was then validated using an independent Japanese data set of 100 patients.

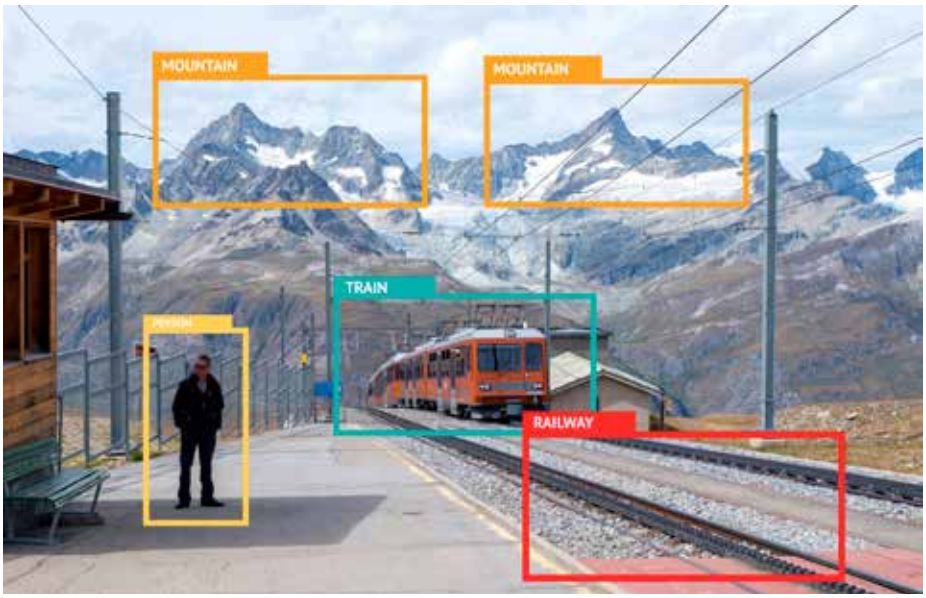
Identifying the underlying mechanisms of type 1 diabetes at the cellular level

Laura Elo studies patient data to search for different biomarkers that can predict disease onset or treatment responses. A biomarker is a factor or characteristic that indicates a change in biological status, for example, in genes or proteins. In Finland, researchers have aimed at determining the underlying mechanisms of type 1 diabetes for a long time. Type 1 diabetes is caused by the destruction of insulin-producing cells. The pancreas does not produce the insulin hormone needed by the body, thereby causing blood sugar to rise.

"Finland has the highest incidence of type 1 diabetes in the world relative to the size of the population. Both genetic and environmental factors play a role in the development of the disease. We look for biomarkers that could predict the development of the disease as early as possible."

Because Finland has the highest levels of type 1 diabetes relative to the population in the world, diabetes research here is also significant. As early as in 1994, the ambitious and extensive research project DIPP (Diabetes Prediction and Prevention) was started in Finland. Genes that predispose you to type 1 diabetes are being sought in the blood samples of newborns. Children who are found to have a genetic risk to develop diabetes are invited to a follow-up study. Samples are taken every three months and, from age 2 onwards, every six or twelve months. The screening participants include the university hospitals in Turku, Tampere and Oulu.

"The children with a genetic risk of developing type 1 diabetes have been monitored until the age of 15. The goal is to iden-



In machine learning, algorithms can make predictions and apply them by analysing data masses.

tify the factors affecting the onset of the disease at the cellular level even before it can be diagnosed with the current methods."

Laura Elo collaborates with Professor **Riitta Lahesmaa**, whose research group studies leucocytes and aims to understand what factors make cells cause diabetes. In the future, this could lead to preventing the onset of diabetes and curing current patients.

New tools

Going forward, Laura Elo wants to focus on the underlying mechanisms of diseases and the risk factors for falling ill. The statistical modelling of the complex interactions between different factors requires many new methods and measurement technologies developed and tested by researchers.

In addition to statistical modelling, Elo and her team applies different machine learning techniques to create predictive models. The machine is taught to learn the essential factors from the data. For example, the machine can learn to provide binary predictions of the consequences of treating an illness with medication: good response/bad response.

"New tools and methods must be brought as close to the patient as possible. We are constantly thinking what is required so that the model can be used in treating patients. What should be measured and how? Is there anything that could be done better? The model must be sufficiently simple and easy to use in order to end up at a clinic to be used by a doctor in their everyday work. It is important to know how doctors use them."

"The essential thing about this work is that it is interdisciplinary. Just how much more information can be obtained using computational methods than sieving through the data only manually. Computation has become part of medicine."

The Turku Centre for Biotechnology has its own computer cluster whose computing capacity is supplemented by a connection to the ePouta cloud service of the Finnish ELIXIR node CSC.

"The computing capacity and tools provided by ELIXIR facilitate the utilisation of data produced by other organisations. Utilising European data is important, but the data should be standardised. Making data compatible is a job for a large infrastructure."

Ari Turunen

MORE INFORMATION:

Medical Bioinformatics Centre, University of Turku:
http://elolab.utu.fi

Bioinformatics services provided by the Finnish ELIXIR node CSC:
https://research.csc.fi/biosciences

Biotools, a range of bioinformatics tools provided by ELIXIR:
https://www.elixir-europe.org/services/tools/biotools

ELIXIR collaborates with the US-based **GA4GH** (Global Alliance for Genomics and Health) to utilise genomic data.
https://www.ga4gh.org

CSC – IT Center for Science
CSC – The Finnish 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
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 2821 ● 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