VEIL.AI anonymises patient data better and faster than traditional methods, and retains information more effectively. If necessary, the application can also produce synthetic, fully anonymous statistical data, which cannot be traced back to any individual.

Developed by the Institute for Molecular Medicine Finland (FIMM), the application is now available for the ELIXIR infrastructure, alongside which a joint service is now being developed. An organisation managing data can protect it by entering the related metadata into a scalable cloud service. The service disguises any identity attributes, providing researchers with anonymised and, if necessary, synthetic data.

**Aided by artificial intelligence**

The VEIL.AI application employs a model based on artificial intelligence. The application creates a veil that protects the patient’s identity attributes, but can identify relevant data, which it keeps.

“Occasionally, for example when creating machine-learning models, more data is required—and more quickly—than research ethics committees tend to allow. They require justification for each variable, which is against the essence of machine-learning where maximum amount of variables is wanted without, a priori, assuming too much of their impact as far as the best mode is still being sought,” says commercialisation expert Tuomo Pentikäinen.

This is why, according to Pentikäinen, it makes sense to use synthetic data in the early stages of modelling, as this is what the VEIL.AI method can produce.

This means data which is totally detached from the people it was derived from, but which, in terms of the desired variables, behaves like the original data. However, synthetic data is only one type of data we provide. Usually customers want anonymised data.”

VEIL.AI can find variables regarded as sensitive in terms of revealing a person’s identity, and anonymise them automatically.

“In a better and more organised way, the application can perform heavy calculations and operations concerning the calculation of data partitioning and anonymisation metrics.”

It must be possible to protect sensitive patient data, but many traditional anonymisation models also lose important data in the process. Traditionally patient data has been protected by partitioning and generalising identity attributes within the data. Anonymisation studies how variables divide/partition data into different groups. Then each group
is examined separately, and if some variables are too easy to identify, they are coarsened. Generalisation means that a person’s age can be rounded off by a few years and a professional title can be changed from, say ‘nurse’ to ‘health care professional’.

“So any variables that are too easy to identify are generalised to a sufficiently general level, or perhaps even removed completely. When processing health data, deletion may have to be used quite often if a variable is unique or too easily identifiable,” says Pentikäinen.

This means that generalisation can result in the loss of important patient data. “This tends to happen when an interesting phenomenon (such as an illness) is relatively rare and fairly evenly distributed across the data set. When such data is divided into partitions for anonymisation, it is common for the phenomenon of interest to be even rarer in each partition. In cases like this, traditional methods commonly interpret the interesting data as “outliers” in each partition, and therefore remove it. This is stupid, because with a better chosen strategy, the phenomenon of interest could have been included in the partitions, while retaining important information more effectively.”

Timo Miettinen, Information Systems Manager at the Institute for Molecular Medicine Finland, has an example: a patient with a rare type of breast cancer. Creating data that is too coarse can lead to the disappearance of data on the rare type of disease, because there are so few such patients in the data set.

“A breast cancer patient has one diagnosis, but her genetic profile indicates that she has a rare type of breast cancer. There may be a few cases like this per hospital, meaning that they may be classified as outliers and deleted. But this does not apply to the entire population. If a better view could have been gained of the big picture, this outlier would not have been deleted.”

Timo Miettinen has been involved for a long time in designing information systems that make use of and protect clinical data. Miettinen and his team have developed the VEILAI application, which is about to become commercially available. This micro service was created due to GDPR.

Each biobank in Finland has its own code register. The code register consists of personal identity codes and a synonym table, used to create an identifier—that is, a pseudonym—for each person.

“Certain things are difficult to change, such as height, eye colour and place of birth. They are identifiable through statistical methods. The same applies to a person’s medical history,” says Miettinen.

“We make two promises. First, we promise scalability and better performance. We are able to make use of continuously updated data from a number of sources. We can anonymise this effectively and securely. Our second promise is to try to minimise data loss. The application takes account of the data content, while fulfilling the anonymisation criteria,” says Miettinen.

Scalable cloud service

The VEILAI application uses a neural network that can be thousands of times faster than conventional methods.

“Our system enables safer data distribution, because once the neural network has been taught what to do, each owner of confidential data can anonymise data before passing it onto its partners. Our method also produces better data, because we can test a
large number of different data partitioning strategies and pick the one that results in the smallest loss of information, while nevertheless achieving the desired level of anonymity," says Pentikäinen.

And crucially, in terms of data security, the VEIL.AI application does not store patient data in a new location. "We do not want to manage data. Instead, data is streamed through our service, during which we anonymise it and return it immediately to the customer," says Tuomo Pentikäinen.

"We offer a scalable cloud service. Through the user interface, we can enter the necessary data dictionary and teach the algorithm to create the data anonymisation model using example material. The algorithm will learn to process the data, and if more data is added, it is streamed through the cloud service and anonymised," says Timo Miettinen. This means that organisations no longer have to share any of their sensitive data with anyone. Data arrives anonymised from the cloud service for research purposes.

The analysis of various pseudo identifiers requires plenty of processing power, which has been obtained from the ELIXIR infrastructure.

Ari Turunen

VEIL.AI is an artificial intelligence application that can anonymise various types of data: telemetric, sensor and continuously updated data, among others. If necessary, it will generate synthetic data from a chosen data set.