AI model predicts disease risks decades in advance

Scientists from the European Molecular Biology Laboratory (EMBL) and the German Cancer Research Center (DKFZ) have developed an AI model that assesses the long-term individual risk for more than 1,000 diseases. The model, which was trained and tested using anonymized medical data from the UK and Denmark, can predict health events over a period of more than a decade. The model, presented in the journal Nature, is not yet ready for clinical use, but it already opens up new possibilities for developing health strategies.

Can your personal medical history be used to predict the health problems you might face in the next two decades? Researchers from EMBL, DKFZ, and the University of Copenhagen have now shown that this is possible. They have developed a generative AI model that uses large-scale health records to estimate how human health may change over time. It can forecast the risk and timing of over 1,000 diseases, and predict health outcomes over a decade in advance.

This new generative AI model was custom-built using algorithmic concepts similar to those used in large language models (LLMs). It was trained on anonymised patient data from 400,000 participants from the UK Biobank. Researchers also successfully tested the model using data from 1.9 million patients in the Danish National Patient Registry. This approach is one of the most comprehensive demonstrations to date of how generative AI can model human disease progression at scale and tested on data from two entirely separated healthcare systems.

“Our AI model is a proof of concept, showing that it’s possible to learn many of our long-term health patterns and use this information to generate meaningful predictions,” said Ewan Birney, EMBL. “By modelling how illnesses develop over time, we can start to explore when certain risks emerge and how best to plan early interventions. It’s a big step towards more personalised and preventive approaches to healthcare.”

The "grammar" of health data

„Just as large language models can learn the structure of sentences, this AI model learns the "grammar" of health data to model medical histories as sequences of events unfolding over time,“ explains Moritz Gerstung, DKFZ. These events include medical diagnoses or lifestyle factors such as smoking. The model learns to forecast disease risk from the order in which such events happen and how much time passes between these events.

“Medical events often follow predictable patterns,” said Tom Fitzgerald, European Bioinformatics Institute (EMBL-EBI). “Our AI model learns those patterns and can forecast future health outcomes. It gives us a way to explore what might happen based on a person’s medical history and other key factors. Crucially, this is not a certainty, but an estimate of the potential risks.”

The model is suitable for various diseases, especially those with clear and consistent progression patterns, such as diabetes, heart attacks, or septicaemia, which is a type of blood poisoning. However, it is less reliable for diagnoses such as infectious diseases, which depend on unpredictable life events, or very rare diseases.

Probabilities, not certainties

Like weather forecasts, this new AI model provides probabilities, not certainties. It doesn’t predict exactly what will happen to an individual, but it offers well-calibrated estimates of how likely certain conditions are to occur over a given period. For example, the chance of developing heart disease within the next year. These risks are expressed as rates over time, similar to forecasting a 70% chance of rain tomorrow.

Some outcomes, like the risk of hospitalisation after a major event – for example a heart attack – can be forecast with high confidence, while others remain more uncertain. Similarly, forecasts over a shorter period of time have higher accuracy than long-range ones.

Heart attack as an example

The risk of heart attack calculated by the AI model for men aged between 60 and 65 varies between a probability of 4 per 10,000/year and around 100 per 10,000/year, depending on previous diagnoses and the men's lifestyle. Women have a lower average risk of heart attack, but a similarly wide range.

In addition, the risk of heart attack increases with age in both men and women. A systematic evaluation of these calculated risks in different age and gender groups shows that they correspond well with the number of cases observed in a subset of the UK Biobank cohort that was not used to train the model.

The model is calibrated to produce accurate population-level risk estimates, forecasting how often certain conditions occur within groups of people. However, like any AI model, it has limitations. For example, because the model's training data from the UK Biobank comes primarily from individuals aged 40–60, it means childhood and adolescent health events are underrepresented. The model also contains demographic biases due to gaps in the training data, including the underrepresentation of certain ethnic groups.

While the model isn’t ready for clinical use, it could already help researchers:

- Understand how diseases develop and progress over time
- Explore how lifestyle and past illnesses affect long-term disease risk
- Simulate health outcomes using artificial patient data, in situations where real-world data are difficult to obtain or access

In the future, AI tools such as the one described here trained on more representative datasets could assist clinicians in identifying high-risk patients early. With ageing populations and rising rates of chronic illness, being able to forecast future health needs could help healthcare systems plan better and allocate resources more efficiently. But much more testing, consultations, and robust regulatory frameworks are needed before AI models can be deployed in a clinical setting.

“This is the beginning of a new way to understand human health and disease progression,” said Moritz Gerstung, DKFZ. “Generative models such as ours could one day help personalise care and anticipate healthcare needs at scale. By learning from large populations, these models offer a powerful lens into how diseases unfold, and could eventually support earlier, more tailored interventions.”

This AI model was trained using anonymised health data under strict ethical rules. UK Biobank participants gave informed consent, and Danish data were accessed in accordance with national regulations that require the data to remain within Denmark. Researchers used secure, virtual systems to analyse the data without moving them across borders. These safeguards help ensure that AI models are developed and used in ways that respect privacy and uphold ethical standards.

This work was funded by EMBL member state contributions, DKFZ funds and Novo Nordisk Foundation grant.

Publication:
Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Mortensen, Ewan Birney, Tom Fitzgerald & Moritz Gerstung:
Learning the natural history of human disease with generative transformers, Nature 2025
DOI: 10.1038/s41586-025-09529-3

Press release

17-Sept-2025
Source: German Cancer Research Center (DKFZ)

Further information

- German Cancer Research Center (DKFZ)