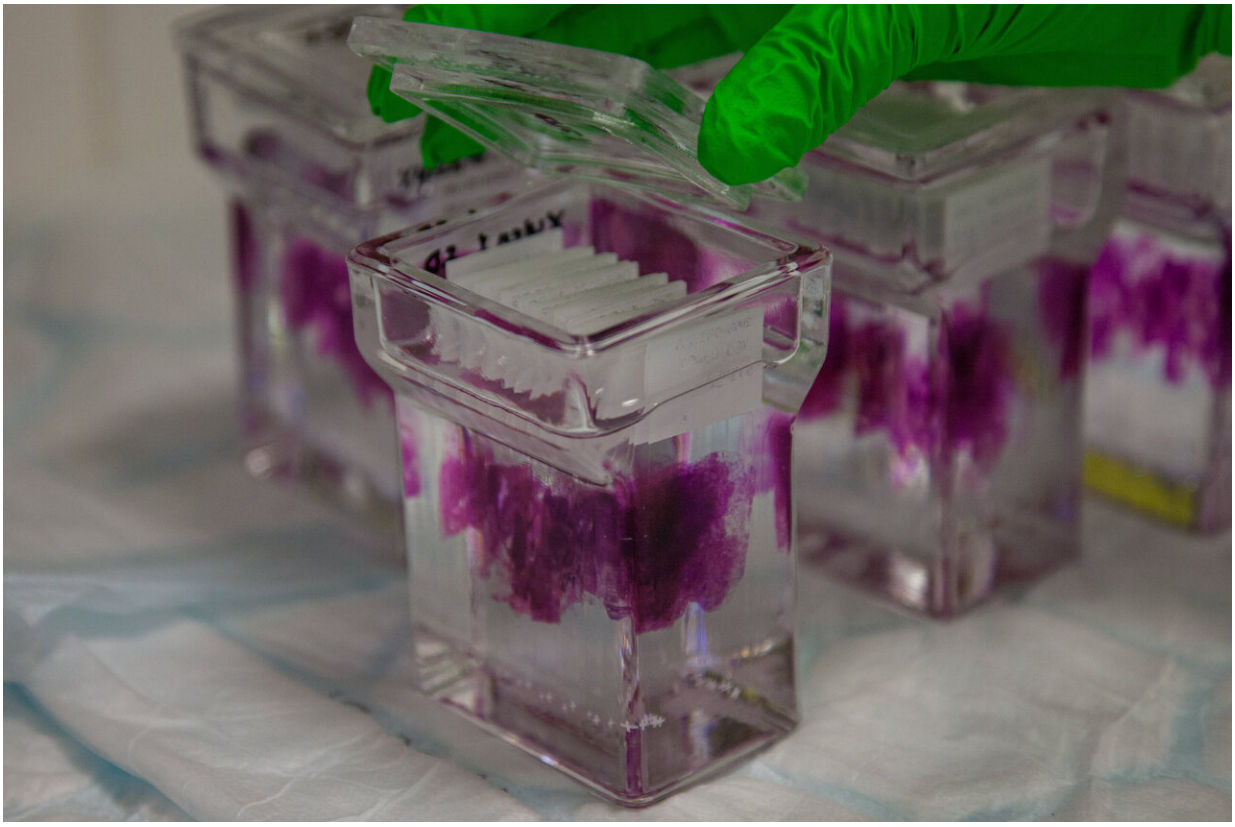
Designing and evaluating medical deep learning systems

February 5 2021



Designing deep learning studies in cancer diagnostics. Credit: Oslo University Hospital

Can better design of deep learning studies lead to the faster transformation of medical practices? According to the authors of

"Designing deep learning studies in cancer diagnostics," published in *Nature Reviews Cancer*'s latest issue, the answer is yes.

"We propose several protocol items that should be defined before evaluating the external cohort" says first author Andreas Kleppe at the Institute for Cancer Diagnostics and Informatics at Oslo University Hospital."

"In this way, the evaluation becomes rigorous and more reliable. Such evaluations would make it much clearer which systems are likely to work well in clinical practice, and these systems should be further assessed in phase III randomized clinical trials."

Slow implementation is partly a natural consequence of the time needed to evaluate and adapt systems affecting patient treatment. However, many studies assessing well-functioning systems are at high risk of bias.

According to Kleppe, even among the seemingly best studies that evaluate external cohorts, few predefine the primary analysis. Adaptations of the deep learning system, patient selection or analysis methodology can make the results presented over-optimistic.

The frequent lack of stringent evaluation of external data is of particular concern. Some systems are developed or evaluated on too narrow or inappropriate data for the intended medical setting. The lack of a well-established sequence of evaluation steps for converting promising prototypes into properly evaluated medical systems limits deep learning systems' medical utilization.

Millions of adjustable parameters

Deep learning facilitates utilization of large data sets through direct learning of correlations between raw input data and target output,

providing systems that may use intricate structures in high-dimensional input data to model the association with the target output accurately. Whereas supervised machine learning techniques traditionally utilized carefully selected representations of the input data to predict the target output, modern deep learning techniques use highly flexible artificial neural networks to correlate input data directly to the target outputs.

The relations learnt by such direct correlation will often be true but may sometimes be spurious phenomena exclusive to the data utilized for learning. The millions of adjustable parameters make deep neural networks capable of performing correctly in training sets even when the target outputs are randomly generated and, therefore, utterly meaningless.

Design and evaluation challenges

The high capacity of neural networks induces severe challenges for designing and developing deep learning systems and validating their performance in the intended medical setting. An adequate clinical performance will only be possible if the system has good generalisability to subjects not included in the training data.

The design challenges involve selecting appropriate training data, such as representativeness of the target population. It also includes modeling questions such as how the variation of training data may be artificially increased without jeopardizing the relationship between input data and target outputs in the training data.

The validation challenge includes verifying that the system generalizes well. For example, does it perform satisfactorily when evaluated on relevant patient populations at new locations and when input data are obtained using differing laboratory procedures or alternative equipment? Moreover, deep learning systems are typically developed iteratively, with

repeated testing and various selection processes that may bias results. Similar selection issues have been recognized as a general concern for the medical literature for many years.

Thus, when selecting design and validation processes for diagnostic deep learning systems, one should focus on the generalization challenges and prevent more classical pitfalls in data analysis.

"To achieve good performance for new patients, it is crucial to use various training data. Natural variation is always essential, but so is introducing artificial variation. These types of variation complement each other and facilitate good generalisability," says Kleppe.

More information: Andreas Kleppe et al. Designing deep learning studies in cancer diagnostics, *Nature Reviews Cancer* (2021). DOI: 10.1038/s41568-020-00327-9

Provided by Oslo University Hospital