6. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; 577(7788): 89–94.

7. Morin O, Vallières M, Braunstein S, et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nat Cancer* 2021; 2(7): 709–722.

8. Chen A, Montesano G, Lu R, et al. Visual field endpoints for neuroprotective trials: a case for AI driven patient enrichment. *Am J Ophthalmol* 2022; 243: 118–124.

9. Urbina F, Puhl AC and Ekins S. Recent advances in drug repurposing using machine learning. *Curr Opin Chem Biol* 2021; 65: 74–84.

10. Romeo-Guitart D, Forés J, Herrando-Grabulosa M, et al. Neuroprotective drug for nerve trauma revealed using artificial intelligence. *Sci Rep* 2018; 8(1): 1879–1879.

Artificial intelligence will change MS care within the next 10 years: No

Maarten De Vos and Jeroen Van Schependom

In this short manuscript, we argue why artificial intelligence (AI) will not change multiple sclerosis (MS) care and definitely not within the next 10 years. We specifically wonder how AI would enable something that *human* intelligence cannot achieve. Without going into a rhetorical discussion about what 'intelligence' actually is, current approaches of AI would rather refer to automated intelligence. Current AI has shown promise at (quickly) automating what human experts can do, for example, sifting through enormous amounts of data, but it has not yet been able to generate novel insights itself. At best, it might extract some insights from a large data set when experts have provided accurate labels. Those insights emerge then from the underlying human-curated data rather than from the power of a specific AI algorithm. As such, similar insights could have been obtained using more traditional methods such as machine-learning (available for over 30 years) or statistics.

Let us take a detailed look at what is needed for a hyped use-case of AI: extraction of novel biomarkers for improved prognosis. Very similar challenges arise in other use-cases of AI, from personalised medicine to drug discovery. A specific AI algorithm is trained on a specific data set with specific labels in such a way that it can reproduce the learned behaviour on unseen data. However, predicting labels with AI for unseen data will be close to expert labels *only* when properties of training and unseen data are similar. This leads to the key challenge: on which data to train the algorithm.

How large should the data set be?

In a practical study, Marek et al. demonstrated that a reliable estimation of a basic correlation coefficient requires thousands of magnetic resonance (MR) images in order to report non-inflated estimates of effect sizes[1] and one can expect that more complex algorithms would require even larger data sizes. In general, AI experts have not been able to derive theoretical guarantees on data set size. Even without knowing the needed size, collecting a sufficiently large data set requires that different laboratories across different countries agree on the 'ideal' set of features (or modalities), which should be collected in a consistent way. Collecting data from multiple centres induces potential data mismatches with respect to patient populations, therapy (history) and sensor manufacturers.

How do we collect such a data set?

Moreover, the data set should be consistent, and follow-up routines cannot be changed during data collection. This is a major hurdle in a rapidly changing clinical context with new drugs/therapies/regulations regularly being introduced.[2] In a different clinical context, Chen et al. observed that a small but recent training sample (1 month, 1800 patients) outperformed a larger (12 months, >10,000 patients) sample and estimated clinical data half-life in the specific context of emergency admissions to be about 4 months.[3] A continuously learning system, which constantly updates and provides updated predictions, is no solution, as such a system cannot be continuously tested in clinical trials (see below).

Correspondence to:
J Van Schependom
AIMS, Center for
Neurosciences, Vrije
Universiteit Brussel, 1050
Brussels, Belgium.
**Jeroen.van.schependom@
vub.be**

Maarten De Vos
Departments of Electrical
Engineering (ESAT)
and Development &
Regeneration, KU Leuven,
Leuven, Belgium

Jeroen Van Schependom
AIMS, Center for
Neurosciences, Vrije
Universiteit Brussel,
Brussels, Belgium;
Department of Electronics
and Informatics (ETRO),
Vrije Universiteit Brussel,
Brussels, Belgium

How can we ensure GDPR compliance?

Collecting and analysing personal data also raises ethical concerns. Current General Data Protection Regulation (GDPR) regulations strictly regulate and complicate sharing of large data sets in order to protect each individual's privacy. While this may potentially be addressed through the concept of federated learning, major hurdles in communication overhead and security issues (see, for example, Mammen's study[4]) have to be addressed.

How can we ensure generalisability?

If we assume we have trained an AI algorithm on a large, high-quality, unbiased data set which provides a prediction on disease evolution, the key question is generalisation performance: will the algorithm predict accurate labels (i.e. prognosis of the patient) on unseen data? While good practice in AI assesses this by splitting data in training, validation and test set, one does not know the properties of the unseen data. Moreover, researchers often train multiple models on an existing data set and report the best-performing one on the test set, leading to overly optimistic results. Similarly, data leakage, that is, properties from the test set leaking into the training data, can be subtle and fuel a new reproducibility crisis in science.[5]

How could we validate AI?

If we expect AI to reveal new insights in disease prognosis, careful validation to the standards for novel treatments/care is needed. So, a proper randomised clinical trial would need to be implemented in which a head-to-head comparison is made between treatment recommendations of the AI algorithm and treatment recommendations by the neurologist. By design, such a trial would require a sufficiently long follow-up time (e.g. 5 years) to be able to prove/disprove long-term potential.

Would you trust a black-box?

Finally, while highly unlikely, assume a fully automated AI system would claim to predict prognosis accurately. Would patients/caregivers follow the recommendation of a black-box AI system? The often-claimed key to trust the system is the ability to explain what patient-specific factors contributed to a certain decision. Unfortunately, this typically assumes a simpler model and thus a reduced accuracy. Novel methods, such as intrinsically explainable methods (see e.g. interpretable boosting models)[6] are being developed but still in their infancy and not ready to be deployed in an actual clinical environment.

Conclusion: data is key

While we have highlighted the reasons why we think AI will not change MS care in the next 10 years, we also want to stress that novel multimodal biomarkers can be uncovered in high-quality data sets that do not require thousands of patients. Several recent papers highlight the importance of these smaller, high-quality and more dynamic data sets.[1,3,7] This again confirms the critical innovation is in how and which data we can collect and much less in which AI model we will use to mine the data.

Declaration of Conflicting Interests

Funding

References

1. Marek S, Tervo-Clemmens B, Calabro FJ, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* 2022; 603(7902): 654–660.

2. Piehl F. A changing treatment landscape for multiple sclerosis: Challenges and opportunities. *J Intern Med* 2014; 275: 364–381.

3. Chen JH, Alagappan M, Goldstein MK, et al. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform* 2017; 102: 71–79.

4. Mammen PM. Federated learning: Opportunities and challenges, 2021, http://arxiv.org/abs/2101.05428

5. Gibney E. Is AI fuelling a reproducibility crisis in science. *Nature* 2022; 608: 250–251.

6. Nori H, Jenkins S, Koch P, et al. InterpretML: A unified framework for machine learning interpretability, 2019, http://arxiv.org/abs/1909.09223

7. Strickland E. Andrew NG: Unbiggen AI. IEEE Spectrum, 9 February 2022, https://spectrum.ieee.org/andrew-ng-data-centric-ai