Artificial intelligence will change MS care within the next 10 years

De Vos, Maarten; Van Schependom, Jeroen

Published in:
Multiple Sclerosis

DOI:
10.1177/13524585221125376

Publication date:
2022

License:
Unspecified

Document Version:
Accepted author manuscript

Artificial Intelligence will change MS care within the next ten years: NO

In this short manuscript, we argue why artificial intelligence (AI) will not change MS care, and definitely not within the next ten years. We specifically wonder how *artificial* intelligence 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, e.g. 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 dataset 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* dataset 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 dataset be?

In a practical study, Marek et al demonstrated that a reliable estimation of a basic correlation coefficient requires thousands of 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 dataset size. Even without knowing the needed size, collecting a sufficiently large dataset requires that different labs 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 dataset?

Moreover, the dataset 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, >10000 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).

How can we ensure GDPR compliance?

Collecting and analysing personal data also raises ethical concerns. Current GDPR regulations strictly regulate and complicate sharing of large datasets 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 eg. [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 dataset 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 dataset and report

the best performing one on the test set, leading to overly optimistic results. Similarly, data leakage, i.e. 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. So, a proper randomized 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 ten years, we also want to stress that novel multimodal biomarkers can be uncovered in high-quality datasets that do not require 1000s of patients. Several recent papers highlight the importance of these smaller, high quality, and more dynamic datasets [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
The authors have no conflicting interests to declare with respect to the research, authorship and/or publication of this article.

Funding
The authors received no financial support for the research, authorship and/or publication of this article.

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**(4):364–381.
3. **Chen JH**, **Alagappan M**, **Goldstein MK**, **Asch SM**, **Altman RB**. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int. J. Med. Inform.* 2017; **102**:71–79. Available at: http://dx.doi.org/10.1016/j.ijmedinf.2017.03.006.
4. **Mammen PM**. Federated Learning: Opportunities and Challenges. 2021. Available at: 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**, **Caruana R**. InterpretML: A Unified Framework for Machine Learning Interpretability. 2019:1–8. Available at: http://arxiv.org/abs/1909.09223.
7. **Strickland E**. Andrew NG: Unbiggen AI. *IEEE Spectr.* 2022.