The Need for Uncertainty Quantification in Machine-assisted Medical Decision Making

Edmon Begoli[1,*], **Tanmoy Bhattacharya**[2,], **and Dimitri Kusnezov**[3,]

[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Los Alamos National Laboratory, Los Alamos, NM, USA
[3]The Department of Energy, Washington, DC, USA
[*]begolie@ornl.gov

ABSTRACT

Medicine, even from the earliest days of Artificial Intelligence (AI) research, has been one of the most inspiring and promising domains for the application of AI-based approaches. Equally, it has been one of the more challenging areas to see an effective adoption. There are many reasons for this, primarily the reluctance to delegate to machine intelligence a function of life-critical and patient-safety-related decision making. To address some of these challenges, medical AI, especially in its modern data-rich Deep Learning guise, needs to develop a principled and formal Uncertainty Quantification (UQ) discipline, just as we have seen in fields such as nuclear stockpile stewardship or risk management. The data-rich world of AI-based learning, and the frequent absence of a well-understood underlying theory poses its own unique challenges to straightforward adoption of UQ. These challenges, while not trivial, also present significant new research opportunities for the development of new theoretical approaches, and for the practical applications of UQ in the area of machine-assisted medical decision making. Understanding prediction system structure and defensibly quantifying uncertainty is possible, and, if done, can significantly benefit both research and practical applications of AI in this critical domain.[1]

Decisions deeply informed through computer modeling, throughout its seventy-years history, have shaped both our paradigm of model-based prediction and supercomputer architectures from transistors to full systems. Starting from well-defined questions, analytic models are sewn together over many length and time scales to yield numerical answers. But numerical results, without a measure of their veracity do not provide the trust needed to inform decisions. Hence, fields of activity on prediction, validation against available data, and how to test algorithms, models, and sensitivities, feed into overall measures of confidence captured under Uncertainty Quantification (or UQ). To achieve this confidence, UQ extends the traditional discipline of statistical error analysis to also capture uncertainties due to possibly incomplete, inaccurate, and contradictory input data, missing and undetected mechanisms and dependencies, expert judgment, and variations between reasonable model forms and modeling strategies. Advancements in UQ now provide measures of confidence necessary to inform national or international security decisions. A notable example is the US support of a nuclear test moratorium since 1992, whereby we annually provide detailed measures of confidence in the safety, security, and performance of the nuclear stockpile—-guaranteed through virtual testing[1,2].

1 UQ in Model-based Critical Decision Making

In model-based prediction, we first understand and define the questions we are posing and then define models to answer them. Not so with data rich problems, where often neither the questions nor the underlying models are known. In this case, artificial intelligence (AI) based methods, from novel hardware to machine

---

learning (ML) techniques, seek to define the effective models that characterize emergent features in data. And that data is often complex, multimodal, discordant, noisy and incomplete.

UQ today underpins many decision processes in nuclear security, our risk management and associated investments, which can be at the scale of billions of dollars. Predictions without UQ are neither predictions nor actionable. The data-rich world of ML, especially the powerful Deep Learning (DL) models, poses parallel challenges. To develop consequential decision-support from "learned" models built on complex data sets, there is an important need to co-develop UQ for this domain. Ultimately it is in the merging of these two distinct worlds—model and data based—that a future path for prediction lies. To get there, an immediate need is UQ for AI based approaches.

In this article, we first discuss some of the current role of DL in clinical decision making. We then describe the possible place and the role for UQ, the challenges this brings forth, how these relate to previous uses of UQ, and fruitful areas of research that we can foresee. We end by summarizing that with a principled approach one can reap the benefits of data-driven approaches without sacrificing our ability to make and defend our clinical decisions.

2 Machine-assisted Clinical Decision Making and Research

Although AI-based research has long played an important role in medicine, it has, nevertheless, been one of the more challenging areas for AI to see an effective adoption. There are many reasons for this, not only social or cultural, but also due to unfamiliar interfaces to decision-making needs and a reluctance to delegate to machine intelligence a function of life-critical and patient- safety-related decision making. On the other hand, the opportunities for the applications of AI in medicine are broad, and, in some areas, possibly transformational. They range from noncontroversial and fundamental applications such as image classification and information extraction, to much more complex, challenging, and possibly high-impact applications such as medical and therapeutic discoveries, outcome predictions, treatment personalization and optimizations, targeted therapies, and possibly far-reaching basic science discoveries.

These are the areas where AI could potentially have radical impacts, but also where errors can have catastrophic consequences. Automated systems adoption, especially systems not analyzable in terms of known causal connections, will require principled and formal UQ to play a transformative role, just as we have seen in the nuclear security domain. UQ captures our pragmatic approaches to ascribing confidence in predictions from some of the most complex simulations done today.

To analyze the situation, we can think broadly of the two streams in empirical sciences—those that (i) use data to derive partial theories or "generalizable/transferable knowledge" that provide understanding and use such knowledge to intervene; or those that (ii) use data to build models that are specific to the problem. The latter do not necessarily provide "understanding", but may use complex correlations in the data to directly make actionable projections. Historically, medicine was squarely in the second category, i.e., it was mainly an empirical science through much of its history, with the rise of statistical interpretations only in the 1950s through the introduction of randomized clinical trials. Even with statistically orientated, and clinical trials, most partial theories in medicine still explain an extremely small fraction of the observed phenomena and variations[3,4]. Even though fully mechanistic models are unlikely to be the first avenues of progress, use of scientific insights and attempts at a cohesive framework incorporating the major clinical predictors are likely to be increasingly useful as predictive models are able to efficiently summarize more complex correlations in the data.

2.1 The Current Roles and Applications of AI

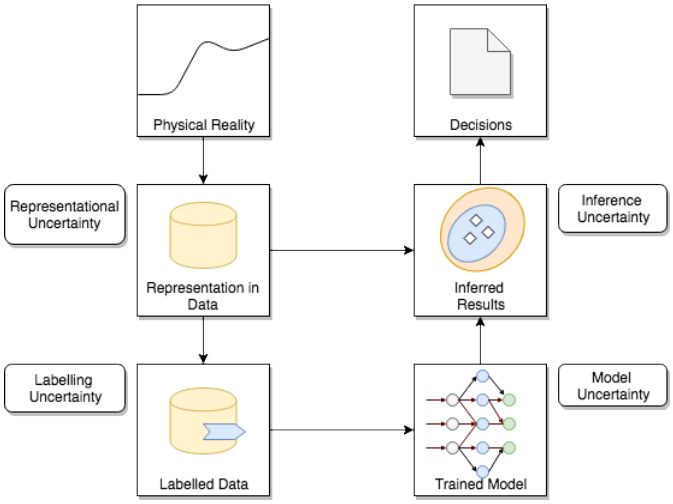
The role of AI in medicine ranges from the well-established tasks of recognition of medical conditions and symptoms with human-like or superhuman accuracy from visual sources, to more novel applications such as

outcome prediction, augmentated cognition, and ultimately guiding medical discoveries and therapy development.

Recently, approaches based on DL have had the most significant impact in the area requiring interpretation of medical images, as DL-structured neural networks are particularly suitable for recognition of visually manifested conditions such as changes in tissue, lesions and growth, etc. The applications of DL techniques based on transfer learning have reported performance comparable to that of the human experts[5], [6] or better[7]. [8] Additionally, DL methods have been used in the predictive scenarios related to quality of care, and clinical outcomes where large neural networks were used as function estimators in place of classical predictive models, with reported performance better than the state-of-the-art, classical model approaches[9]. [10,11]

Finally, there is a growing application of AI techniques in discovery-oriented biomedical subdisciplines. Some are in more applied areas such as drug discovery, and some are in more fundamental science areas such as the study of chemical reactions[1,2] and assistance in the exploration and discovery of the molecular characteristics of medical phenomena from the available data using deep learning and other AI methods[1,3,14]. In most of the presented cases, the applications of AI are based on the DL neural networks, trained on a very large number of labeled data sets, and their learning tuned with the large number of hyperparameters. The most commonly applied neural network architectures are Convolutional Neural Networks (CNNs) for the analysis of images, Recurrent Neural Networks (RNNs) for analyzing time series and prediction, and sequence recognizers (e.g. LSTMs[15]) for the analysis of text, though the architecture of the network is itself often the subject of exploration[1,6] Unlike statistical approaches where mathematical models are used to explain variations observed in data, and to propose the margin of errors on inferences, with these recent applications, different learning architectures are combined with a large number of DL network parameters to form universal approximators. These are then "trained" to reconstruct the outcome of some generative function, without an explicit attempt to specify the exact mathematical model behind the process.

Figure 1. Some of the sources of uncertainty in today's DL pipeline.



3 The Role for UQ in DL

To understand the role of UQ in DL, we need to understand the lifecycle of a typical DL process, and how UQ might fit into it. Unlike classical scenarios that start with the formulation of models reflective of physical reality, almost all DL scenarios start with the collection of the potentially relevant, and the most comprehensive data set available for decision-making scenarios (for example, an early detection of the onset of sepsis). Collection and organization of data is often followed, unless data is already labeled, by "labelling" of

the data to mark the phenomena of interest (e.g. patterns of vital signs characteristic of the pre-septic patients). These data are then used for training the DL model to meet some performance goals (e.g. accuracy and precision in prediction of patients with pre-septic clinical features). To enable this, the authors of the DL process first select the most suitable DL architecture for this kind of predictive application, and then train the DL network with the labeled data. This training process is iterative, and involves the optimization of a variety of the learning parameters, which will be "tweaked" until the network is trained to a sufficient level of performance. Next, the trained model is validated against the validation dataset – the dataset that has not been previously "seen" by the network. If the performance of the model meets the desired performance criteria, the model will be deemed potentially usable in intended scenarios (e.g. early onset of sepsis surveillance). Obviously, there are many steps in such scenarios where there are uncertainties that would need to be quantified. The obvious ones are uncertainties related to the (i) collection and selection of the training data and how well it represents and covers the actual medical phenomena, (ii) accuracy and completeness of the labelling of the training data, (iii) selection and understanding of the actual DL model, and its performance bounds and limitations, and (iv) uncertainties related to model's performance against the operational data (clinical inference). While still non-exhaustive, we propose that all of these steps (illustrated in figure 1) would need to be quantified in order to arrive at even crude, overall measures of the uncertainty of the DL-based decision model.

4 Anticipated Challenges

We see at least four overlapping groups of challenges associated with the uncertainty quantification of the data-driven approaches such as DL:

1. *Absence of theory:* Unlike the physical world which is governed by the well-understood laws of physics, the domains where the DL is usually applied, such as medicine, do not have such "hard laws". Although we use compensating mathematical techniques that take certain assumptions in order to account for the random noise, or some other well-known problem in working with the data, we are ultimately operating without the fundamental, underlying mathematical model which we could otherwise use to ground our uncertainties and to bound any assumptions we make.

2. *Absence of causal models:* In addition to the absence of underlying mechanistic theory, one also has to contend with the fact that DL is essentially exploiting correlations in the data, without paying attention to any causal link. This may not seem like a limitation since prediction does not need causal relation. In fact, once a low-dimensional representation is arrived at describing certain correlations (e.g., the difference between cancerous and matched normal cells), it can raise hypotheses that can be tested. The absence of a causal connection, however, means garnering limited conclusions from DL models; furthermore, it is imperative to understand how the training data must be similar to prediction data.

3. *Sensitivity to imperfect data:* As we discussed before, DL learns from data, and often uses subtle multivariate correlations to improve its predictions. Real world data is usually imperfect—typically containing missing elements and errors—and these imperfections have patterns that can confound prediction. Specific UQ methods, therefore, need to be developed to quantify the sensitivity of models to imperfect data.

4. *Computational Expense:* The training of the DL models is computationally expensive, and any further re-computation and re-evaluation of the models, aimed, for example, at the calculation of uncertainty bounds might currently be prohibitively expensive. Fortunately, computing capacity in support of DL is growing exponentially, and techniques are being developed[17] to approximate some of the UQ-relevant calculations.

To note, *ad hoc* solutions, such as sensitivity analysis and study of model variability, have sometimes been employed to mitigate some the problems we outline. A need to systematize a similar situation is what actually led us to develop the formal approach to UQ in US national security sciences. The wider application of DL in the biomedical field now requires an extension of these methods to this emerging field.

5 Needs for New Research

Just as the challenges in applying UQ for DL are significant, the opportunities for new and important research are equally exciting. Even though a review of the ongoing research in this area is beyond the scope of this article, in this section we describe a few major research directions that, we expect, could improve the situation. In the end, it is possible that the entire new field of UQ for DL might need to be developed.

5.1 Quantifying and limiting overfitting

Overfitting, or the problem of a model performing well on the training set, but generalizing poorly for unseen datasets, is one of the fundamental problems of all data-centric methods, and therefore DL.

In classical models, we evaluated models' performance by information criteria that strongly penalized the number of parameters estimated from the data, and strong guarantees against overfitting relied on proving that the assumptions did not allow one to fit random noise. With DL, and the large number of model parameters involved as well as the capacity of DL networks to memorize random noise[18], classical approaches do not work.

The research question in the context of DL is: what is a scheme that informs us about the bounds of overfitting. Some approaches, such as attempts to empirically learn generalizable patterns with insertion of random noise[19,20], or the use of cross-validation to determine the progression of generalizable learning, move us forward in this problem space, while still carrying the problem of overfitting[21]. Despite these advances, further research is needed in the criteria that can be used to provide provable limits on overfitting assuming fair sampling in the training data.

5.2 Understanding DL

Advances in understanding of how DL works internally will allow for a more effective UQ of interpreting DL. This is an active area of research, with a common approach focusing on interpreting the relationship between the input and output of a DL algorithm, and providing an explanation of the results, not only on individual instances, but of general method. In addition, there are ongoing studies that attempt to understand what DL does,[22] and how it learns[23]

5.3 Training DL to provide its own uncertainty estimates

Ultimately, an effective way of addressing some of the mentioned UQ problems might be to re-shape the DL engine itself to provide an uncertainty estimate on its predictions. In other words, instead of trying to analyze a trained DL network, or the training procedure, one can use the characteristics, architecture, and computational capabilities of the DL process to learn to analyze its own uncertainty. We propose this based on the realization that, ultimately, uncertainty in generalization depends on the density of training points in an appropriately defined neighborhood of the prediction target. In high-dimensional problems, typical to the medical setting (images, large numbers of phenotypes, etc.), every point, however, can be isolated in some other dimension, and a density of points makes sense only after irrelevant dimensions, are projected out – this is difficult to do just by analyzing the network from outside. On the other hand, the network itself can be used to study this uncertainty empirically and provide the uncertainty bounds. A particularly fruitful approach seems to be the use of Generative Adversarial Networks (GANs) for detecting out-of-sample cases.[24–26]

6 Summary

Data-driven methods are emerging as the foundations of evidence-based decision making, and the future of data-driven scientific discovery. To fully realize their potential, we need to overcome significant hurdles in

understanding the precision and uncertainty in purely data-driven predictions. Fortunately, there is a unifying structure to this problem in its various guises of complex predictive correlations in large data sets and engineering black boxes. These have been studied in other scientific disciplines and decision-making arena, and we can learn from those. The details of the medical applications and DL networks are, however, significantly different since the theoretical foundations are far less developed and there are deep psychological and sociological implications in delegating to machines decisions that affect the life or health of human beings. Nevertheless, progress in understanding the structure of these predictive systems, merging model and data driven approaches with strongly defensible UQ, and the formalization of the UQ for DL discipline will be needed to make DL and other data-centric tools and methods practically useful. UQ for DL will likely not be simply a set of tools or procedures to apply, but a more complex wrap-up of disparate methods that in total help bound the overall confidence in predictions.

Acknowledgments

This manuscript has been in part co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725.

References

1. Oberkampf, W. L. & Roy, C. J. *Verfication and Validation in Scientific Computing* (Cambridge University Press, 2010).
2. National Research Council. *Evaluation of Quantification of Margins and Uncertainties: Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile* (Washington, DC: The National Academies Press, 2009).
3. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109, 1193–1198 (2012).
4. Choi, J. D. & Lee, J.-S. Interplay between epigenetics and genetics in cancer. *Genomics & informatics* 11, 164–173 (2013).
5. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131 (2018).
6. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115 (2017).
7. Mar, V. & Soyer, H. Artificial intelligence for melanoma diagnosis: How can we deliver on the promise? *Annals of Oncology* (2018).
8. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one* 12, e0174944 (2017).
9. Bychkov, D. *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports* 8, 3395 (2018).
10. Xiao, C., Ma, T., Dieng, A. B., Blei, D. M. & Wang, F. Readmission prediction via deep contextual embedding of clinical concepts. *PloS one* 13, e0195024 (2018).
11. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 1, 18 (2018).
12. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* 3, 434–443 (2017).
13. Hsu, E., Klemm, J., Kerlavage, A., Kusnezov, D. & Kibbe, W. Cancer moonshot data and technology team: Enabling a national learning healthcare system for cancer to unleash the power of data. *Clinical Pharmacology & Therapeutics* 101, 613–615 (2017).
14. Fillon, M. Making sense of the mountains of new cancer data. *JNCI: Journal of the National Cancer Institute* 109 (2017).
15. Geraci, J. *et al.* Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-based mental health* 20, 83–87 (2017).
16. Zhou, Y. *et al.* Resource-efficient neural architect. *CoRR* abs/1806.07912 (2018).
17. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning (2015). arXiv:1506.02142.

18. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *CoRR* abs/1611.03530 (2016). URL http://arxiv.org/abs/1611.03530. 1611.03530.
19. Arpit, D. *et al.* A closer look at memorization in deep networks (2017). arXiv:1706.05394.
20. Zhang, C., Vinyals, O., Munos, R. & Bengio, S. A study on overfitting in deep reinforcement learning. *CoRR* abs/1804.06893 (2018). URL http://arxiv.org/abs/1804.06893. 1804.06893.

21. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11, 2079–2107 (2010).
22. Brahma, P. P., Wu, D. & She, Y. Why deep learning works: A manifold disentanglement perspective. *IEEE Transactions on Neural Networks and Learning Systems* 27, 1997–2008 (2016).
23. Raghu, M., Gilmer, J., Yosinski, J. & Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability (2017). arXiv:1706.05806.
24. Brahma, P. P., Huang, Q. & Wu, D. O. Structured memory based deep model to detect as well as characterize novel inputs. *CoRR* abs/1801.09859 (2018). URL http://arxiv.org/abs/1801.09859. 1801.09859.
25. Yu, Y., Qu, W., Li, N. & Guo, Z. Open-category classification by adversarial sample generation. *CoRR* abs/1705.08722 (2017). URL http://arxiv.org/abs/1705.08722. 1705.08722.
26. Ge, Z., Demyanov, S., Chen, Z. & Garnavi, R. Generative openmax for multi-class open set classification. *CoRR* abs/1707.07418 (2017). URL http://arxiv.org/abs/1707.07418. 1707.07418.