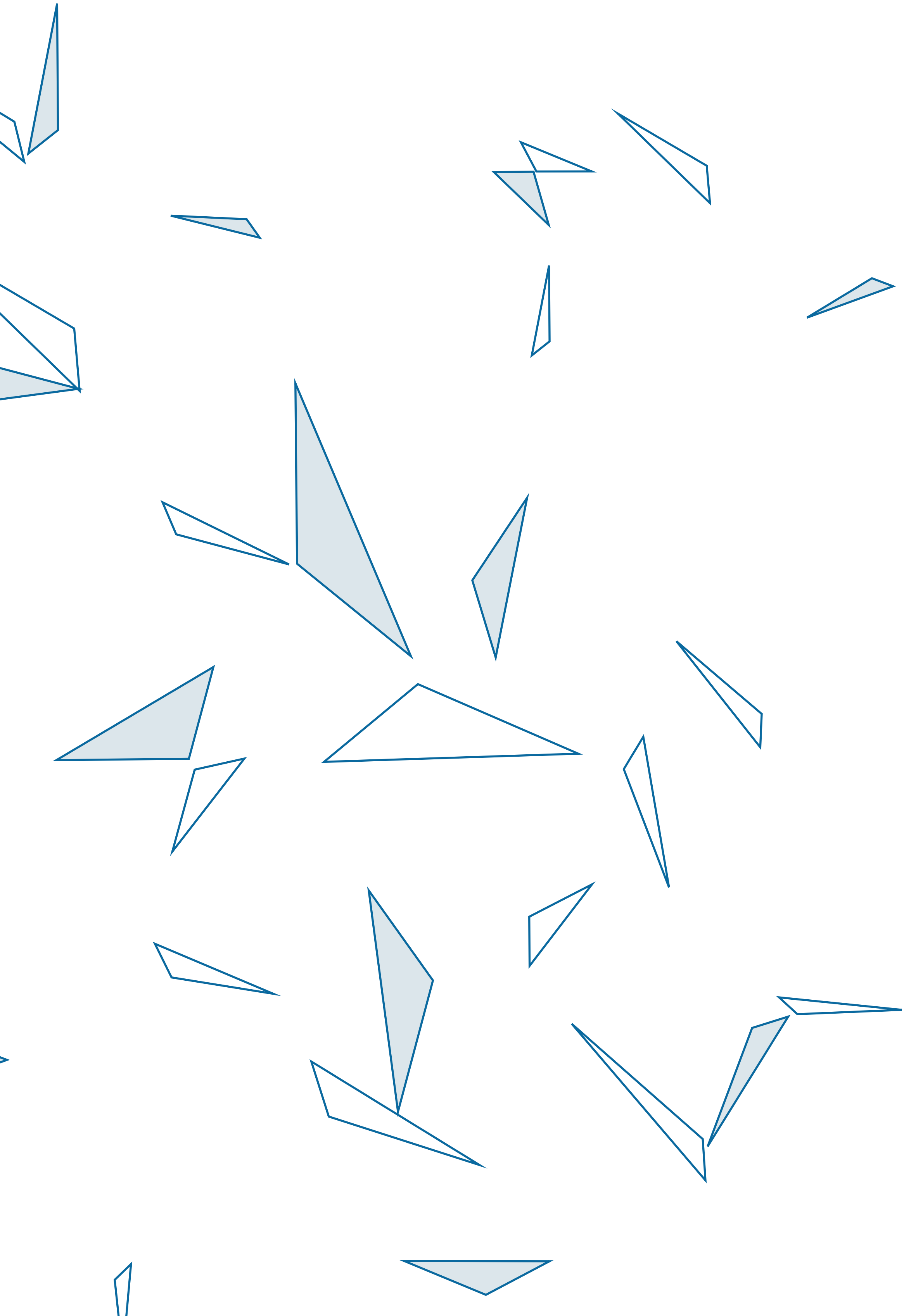
Applied machine learning in neurosurgical oncology

Senders, J.T.



10

General discussion

GENERAL DISCUSSION

Due to the infiltrative nature of the disease, the median expected survival in patients with a malignant brain tumor remains dismal despite improved surgical and adjuvant treatment strategies.[1] The thin line between treatment effectiveness and patient harm underlines the importance of tailoring clinical management to the needs of the individual patient and suggests a strong potential for the emerging field of predictive analytics.

Classical statistics

Throughout the medico-scientific history, numerous analytical techniques have been developed to derive knowledge from experiments and observations to improve day-to-day patient care. Classical statistical methods evaluate the strength of an association between patient characteristics and outcomes within a sample population, with the aim of generalizing these conclusions to the larger population.

Although these statistical techniques have become indispensable for studying treatment efficacy and identifying risk factors, their coefficients remain group-level estimates derived from the total study cohort and do not necessarily apply to the same extent in each individual patient. A clinical trial could demonstrate the efficacy of a novel neurosurgical procedure, as well as the rate of complications observed at the cohort-level. In day-to-day clinical care, however, the question remains to what extent the individual patient would benefit from this treatment and how likely he or she is to experience the dreaded adverse events.

The advent of predictive analytics provides clinicians with the analytical support for personalizing treatment decisions. Regression analysis can compute patient-level predictions of the outcome by adding the population intercept and the slope coefficients pertinent to the individual patient. To develop this model, however, human experts still need to determine which variables to include, identify relevant effect modifiers, and perform data transformations to meet the underlying assumptions. This requires pre-existing human understanding to hypothesize these statistical patterns and substantial effort to define the model properties accordingly. The high level of human interference is feasible for structured data sets with a limited number of clinically interpretable variables and even provides valuable insights into the underlying relationships among variables and outcomes. But how should a human test which variables to select, interactions to include, or transformation to perform if the variables are composed of individual words in a text document, pixels in a picture, genes on a chromosome, or voxels in an MRI scan, let alone specify all these model properties by hand?

Machine learning

This is where machine learning comes into play. In contrast to classical statistics, modern machine learning prioritizes prediction over inference, even if it is achieved at the cost of its interpretability.[2] Compared to regression analysis, the modeling process (e.g., the inclusion of nonlinear associations and interaction terms) occurs rather automatically in many machine learning algorithms. Furthermore, they are less concerned with providing interpretable coefficients but rather oriented towards computing accurate predictions. Because they require less human guidance, these algorithms can model complex patterns automatically, even those that are potentially undetectable or meaningless for humans. Similar to regression analysis, however, classical machine learning algorithms, such as fully connected artificial neural networks, random forest, and support vector machines, are limited to the analysis of structured data (i.e., data in tabular, two-dimensional format in which observations are represented by rows and variables by columns). As a result, a neuroradiologist still has to measure the size of a brain tumor manually and insert this value into a data collection sheet to allow for the construction of classic machine learning models. This poses a significant burden on the clinician or researcher and introduces human subjectivity with regard to the generation and selection of input features. Furthermore, it ignores the potentially relevant hierarchical relationship between individual data points. Voxels close to each other in the scan might have a different, yet relevant relationship compared to voxels far away from each other. This spatial, temporal hierarchy would be missed if the data is shoehorned into a tabular format.

Deep learning has emerged as a family of techniques that were designed to develop models directly from the raw, unstructured data itself.[3] It allows the computer to ingest and analyze high-dimensional data formats (e.g., free text, pictures, MRI scan) and identify meaningful representations within the data. Considering the same neuro-imaging example, nodes in the lower layers of a computer vision model might be susceptible for detecting simple straight lines in the brain MRI, subsequent hidden layers can learn how to detect shapes by recognizing combinations of lines, and the top layers utilize this condensed knowledge to produce clinically meaningful estimates, such as diagnostic classifications, volumetric segmentations, or outcome predictions. This process of condensing high-dimensional data to meaningful features within the model is called feature extraction and allows the raw data to speak for itself.

Instead of engaging into the futile efforts of defining when an algorithm becomes machine or deep learning, they can be considered as an extension of traditional statistical approaches. Machine learning algorithms exist along a continuum, determined by how much is specified by humans and how much is learned by the machine, referred to as the machine learning spectrum.[4] The current thesis describes several studies along the continuum of the machine learning spectrum as it applies to neurosurgical oncology (Figure 1).

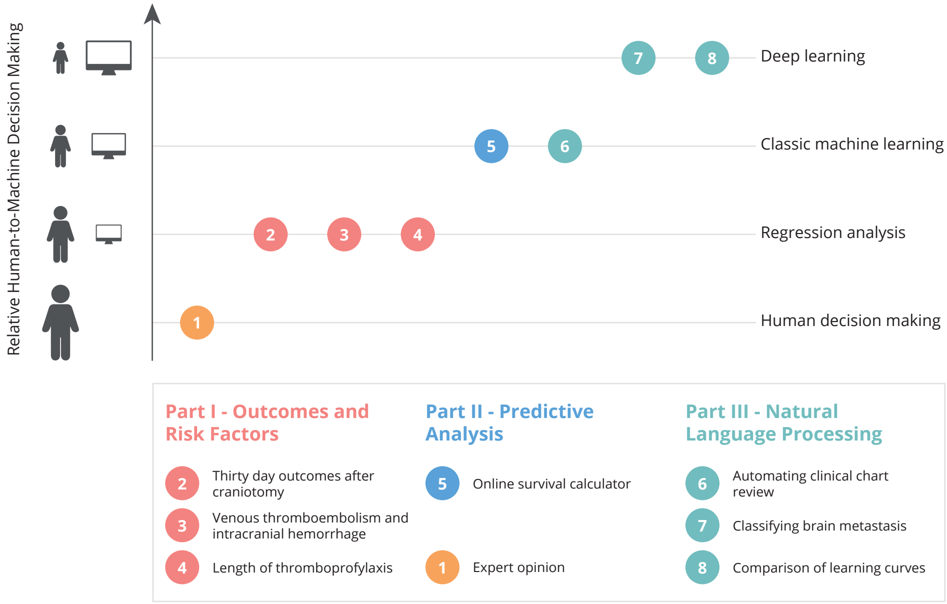

Relative Human-to-Machine Decision Making
7
8
Deep learning
5
6
Classic machine learning
2
3
4
Regression analysis
1
Human decision making
Part I - Outcomes and Risk Factors
2 Thirty day outcomes after craniotomy
3 Venous thromboembolism and intracranial hemorrhage
4 Length of thromboprofylaxis
Part II - Predictive Analysis
5 Online survival calculator
1 Expert opinion
Part III - Natural Language Processing
6 Automating clinical chart review
7 Classifying brain metastasis
8 Comparison of learning curves

FIGURE 1. The machine learning spectrum as it applies to the current thesis. Numbers 2 to 8 correspond to the chapters in the current thesis.

Part I: Outcomes and risk factors in neurosurgical oncology

In **Chapters 2 and 3**, the inferential utility of regression-based algorithms was used to identify risk factors associated with 30-day outcomes in patients operated for a malignant brain tumor. Among patients undergoing craniotomy for a primary malignant brain tumor, 12.9% experienced a major complication within 30 days after surgery, in particular elderly patients and patients with worse functional status or more comorbidity. The increased risk of adverse events should be considered and balanced against the expected survival benefit in this particular patient population. Reoperation and venous thromboembolism were identified as the two most common postoperative major complications, and intracranial hemorrhage as the most common reason for reoperation. These results indicate blood coagulation as a primary challenge in the perioperative management of glioblastoma patient with a careful balance, often deviating in both directions.

In a subsequent in-depth analysis (**Chapter 3**)**,** intracranial hemorrhages occurred predominantly within the first days of surgery, whereas the risk of thrombogenic complications, and pulmonary embolisms in particular, extended beyond the period of hospitalization. The hemorrhagic and thrombogenic risk patterns, which diverge over time, suggest caution with regards to starting anticoagulation shortly after surgery, as well as a potential role for continuing it beyond the period of hospitalization. In a retrospective cohort study investigating this prophylactic strategy (**Chapter 4**), the rate of venous thromboembolism remained nevertheless similar in patients receiving short (i.e., up to discharge) versus prolonged (i.e., 21 days after surgery) thromboprophylaxis. A higher rate of intracranial hemorrhages was even observed in the latter group. Based on these findings, we do not recommend the routine use of prolonged thromboprophylaxis in patients undergoing craniotomy for high-grade glioma.

Part I characterized risk factors of postoperative complications, as well as the safety and efficacy of thromboprophylaxis, in patients undergoing craniotomy for a primary malignant brain tumor. However, the interpretable coefficients to quantify these effects remain group-level estimates and do not necessarily apply to each individual patient to the same extent. After all, the risk of venous thromboembolism in the individual patient can be very different from the cohort's average. Although routine use of prolonged thromboprophylaxis did not significantly reduce the rate of venous thromboembolism at the group-level, this does not preclude selected individual patients to benefit from this strategy. Predictive analytics could help in personalizing clinical decision-making to the characteristics and needs of the individual patient.

Part II: Predictive analytics in neurosurgical oncology

In **Chapter 5**, we developed a model to predict survival in the individual glioblastoma patient. We trained several statistical and machine learning algorithms based on structured demographic, socio-economic, clinical, and radiographic information. The accelerated failure time model demonstrated superior performance in terms of discrimination, calibration, interpretability, predictive applicability, and computational efficiency compared to Cox proportional hazards regression and other machine learning algorithms.

Surgery, and neurosurgery in particular, is characterized by balancing outcome probabilities. In the decision-making process, the surgeon has to weigh the chances of a favorable outcome against the risks of surgery, keeping in mind the natural course of the disease. Large cohort studies allowed us to estimate the expected outcomes in the total population and even differentiate between various risk strata. These strata, however, comprise clusters within the total population, ranked on a single or few cardinal features. As such, the physician still relies on group-level statistics complemented with their own clinical experience. The lack of personalized outcome and risk assessment can result in informed consent procedures that are ambiguous and biased towards the mean (e.g., "Trials have shown a median increase in survival of …", "Generally, X out of 100 will develop …"). Predictive models in contrast intend to quantify the estimated outcomes in the individual patient. As such, a personalized overview of the estimated outcomes can be provided when communicating different surgical strategies with patients and their families. This not only improves the patient selection and surgical decision-making but also enhances the patient's autonomy throughout the decision-making process.

To facilitate its transparency, reproducibility, and utility, we deployed the model developed in **Chapter 5** as an online calculator for survival through a free, publicly available software. This prediction tool provides an online and interactive interface for survival modeling with the potential to inform clinical and personal decision-making in the individual glioblastoma patients. External and prospective validation on heterogenous cohorts from multiple institutions remains necessary, however, to confirm its prognostic value at point-of-care prior to clinical implementation. Furthermore, the online calculator, as well as clinical prediction tools in general, should be considered as dynamic rather than static products developed on the best available evidence available at that point. Continuous model evaluation and optimization remains therefore mandatory to improve its accuracy and precision based on supplementary patient data and novel insights. Currently, we are working on the first model update utilizing the recently published SEER data of glioblastoma patients diagnosed in 2016 as well. This update has improved model performance according to the Harrell's C-index from 0.70 (95%CI 0.70 – 0.70) to 0.73 (95%CI 0.73 – 0.73). In the future, we aim to re-iterate the analysis and further optimize model performance. Collection of information on functional status and molecular markers in the SEER registry could be a valuable first step towards optimizing the model again in the near future.

Part III: Natural language processing in neurosurgical oncology

Part III encompasses the application of machine learning to a higher dimensional problem. Various natural language processing approaches were developed to automate the processing and analysis of narratively written clinical reports. In **Chapter 6**, we have developed a pipeline for automated clinical chart review by analyzing a corpus of free-text radiology reports of brain tumor patients. In this study, we utilized a bag-of-words approach with a classical statistical algorithm known for its strong method of regularization, LASSO regression. The developed pipeline was able to extract 15 distinct radiographic features with high to excellent discriminatory performance (AUC 0.82-0.98). Model performance was correlated with the interrater agreement, which underlines the importance of expert consensus in generating ground truth training labels. However, expert consensus can also be used as a potential indicator for the complexity of the natural language processing task at hand.

In **Chapter 7**, we compared various statistical (logistic regression, LASSO regression), classical machine learning (fully connected artificial neural networks), and deep learning (convolutional neural networks, gated recurrent unit, and long short-term memory) techniques in their ability to classify radiology reports of brain metastases patients into reports that describe solitary versus multiple metastases. Both the LASSO regression and convolutional neural networks model demonstrated to outperform other competing statistical and machine learning models. Although these algorithms are on the opposite ends of the machine learning spectrum, their performance were highly comparable. The LASSO regression model focused merely on the relative frequency of words or word combinations but ignored the order or semantic properties of individual words. In contrast, the deep learning model (i.e., convolutional neural networks) were able to accommodate to higher-level lexical complexity. This sequence-based approach also modeled the order of the words and paragraphs, as well as the semantic relationships among words and thus the statistical properties of a language.

Despite the advantages of modeling these sequential and semantic attributes, the deep learning model in this project did not outperform the less complex LASSO regression model. Perhaps due to its simplicity, LASSO regression demonstrated the most robust performance across different metrics. This implies that the underlying signal for this particular text classification task was found in primitive (i.e., relative word frequencies) rather than complex patterns within the data (i.e., sequential and semantic relationships). By scrutinizing the full complexity of the data, however, deep learning algorithms were computationally inefficient and perhaps prone to overfitting to statistical noise.

In **Chapter 8**, we compared the learning curves of various algorithms in determining the histopathological diagnosis of brain tumor patients based on free-text pathology reports. In this study, we developed a modified version of the generic convolution network model equipped with stronger methods of regularization. The resultant model was able to model the semantic complexity of text documents without overfitting to statistical noise. The number of required training samples to reach the predetermined performance thresholds (an AUC of 0.95 and 0.98) was two to eight times lower for the modified deep learning model, ClinicalTextMiner, compared to regression and conventional deep learning-based architectures. The steep learning curve can be valuable for natural language processing tasks with a limited set of training examples available (e.g., rare diseases and events or institutions with lower patient volumes).

Utilizing natural language processing in healthcare could have profound implications for clinical research and even patient care. Currently, clinical research endeavors are restricted significantly by the need for financial and human resources to gather, process, and analyze clinically generated data. Observational studies are therefore limited to data sets that can be collected by hand, often a mere fraction of the entire population. Yet, their results are generalized to the entire population. The automatic nature could accelerate retrospective chart review to an unprecedented scale, such as the entire population, and allow for the assembly of large, continuously updated clinical registries. The deterministic nature can make data collection less subject to inter- and intra-reviewer inconsistencies but rather based on a consensus label from clinical experts.

The impact of natural language processing in clinical care could even be more profound. Although the bulk of biomedical information is increasing in volume and complexity, the human physician brain that has to comprehend this information is and will remain the same. Information overload, therefore, constitutes a significant problem in the digital age of healthcare and plays a key role in diagnostic errors, near misses and patients' safety, as well as the stress and work satisfaction perceived among healthcare workers.[5,6] Natural language processing algorithms might assist exposing relevant information in a patient's chart without multiple clicks or relieve the administrative burden on clinicians. The process of viewing and entering the clinically most useful data frictionless is essential for clinicians, not just for their convenience, but to spend more time with their patients and provide the best possible care.[7]

Lastly, by extracting and analyzing patient characteristics and outcomes automatically, natural language processing could facilitate a health care system that continuously learns from clinically derived data, thereby narrowing the gap between research and patient care. This resultant collective learning curve can be used to inform and optimize clinical decision-making in the individual patient at point of care.

The machine learning spectrum in neurosurgical oncology

It is the increasing availability of high-dimensional clinical information and computational power that has propelled the use and popularity of algorithms on the high end of the machine learning spectrum (i.e., deep learning). However, these 'black box' algorithms do not constitute the computational panacea to all medico-scientific problems due to the lack of interpretability and need for enormous amounts of data to grasp the full complexity of the data without overfitting.[3] This thesis confirms that placement on the high end of the spectrum does not necessarily imply superiority over other algorithms.

Regression analysis, as demonstrated in **Part I**, and other methods for statistical inference will remain pivotal for clarifying clinically relevant associations at the group-level. It performs well and consistent, even on relatively small data sets. Predictive analytics has the potential to personalize these estimates after collection of sufficient amounts of training data. Even in the predictive realm, however, machine learning does not necessarily outperform classical statistical algorithms, as shown in a recent systematic review as well.[7] In **Part II**, for example, we deployed an algorithm on the low end of the machine learning spectrum (i.e., the accelerated failure time) because of its superior predictive performance and interpretability. In **Part III**, however, we gravitated towards the high end of the machine learning spectrum (i.e., deep learning). In these natural language processing studies, the input consisted of unstructured high-dimensional data, namely free-text clinical reports. Manual specification of the almost infinite number of associations, interactions terms, and data transformations would be virtually impossible and meaningless for humans. Algorithms on the high-end of the machine learning spectrum, on the other hand, allowed for automated analysis of the hierarchical and semantic relationships among words, without the need for manual specification.

Future research

Instead of focusing merely on novel and complex algorithms on the high end of the machine learning spectrum, future research should focus on tailoring the modeling approach to the computational and clinical problem at hand. After all, different problems require different levels of human involvement.

Despite the rapid development of high-performing clinical prediction models, few are actually implemented in the clinical realm. This underlines the importance of shifting our focus from the technical development to the clinical implementation and the ethical challenges that come along with it. At this stage, clinical implementation is not solely dependent on whether we can improve the performance of a given model from 99.0% to 99.5%. It is rather dependent on whether we as a medical society decide to rely our clinical decision-making on the model, while accepting that it is wrong 1% of the time. Future research should therefore focus on developing implementation criteria for high-performing prediction models, considering both the accuracy and clinical consequences of their predictions. Rather than focusing merely on measures of prediction performance, we therefore advocate a multimodal assessment including measures of interpretability as well when developing clinical prediction tools. In addition to implementation criteria, we also advocate the development of mechanisms for continuous performance evaluation and even exit criteria for models that have been clinically implemented. After all, their performance is not a static fact but highly subject to changes in the clinical environment. For example, a sudden, yet undetected change in patient population or data acquisition methods could instantly reduce model performance, and a delay in detecting the deviating performance trends can result in detrimental patient outcomes.

Additionally, we underline the importance of adopting the concept of open source coding in clinical research. Open source coding enhances the reproducibility and transparency of machine learning models developed in medical research. As such, it facilitates the implementation and acceptance in clinical care as well.[8] To allow for external validation, we have deployed the model developed in **Chapter 5** as a publicly accessible, online survival prediction tool for glioblastoma patients. In **Chapters 6, 7, and 8**, we did not deploy the resultant natural language processing models because these models were trained on a text corpus of a single institution, which may be characterized by unique styles and language in their clinical reports. As such, they may not generalize well to text corpora from external institutions or documents written in other languages. Instead, we released the underlying source code which allows for the development, validation, and optimization of similar models in other languages, institutions, patient populations, clinical reports, and outcomes.

In addition to enhancing the transparency of prediction models, improving the computational knowledge among clinicians can reduce a dependency on ‘black-box’ algorithms and shift the doctor-versus-machine paradigm to a doctor-and-machine paradigm. Although optimization of the internal parameters occurs automatically, model fitting only constitutes a single step within the process of model development that largely occurs outside the 'black-box'. For example, the way patients are selected, input features preprocessed, complexities in the data accounted for, outcomes defined, hyperparameters optimized, and model performance evaluated are all specified manually based on clinical expertise and substantially determine the internal and external structure of the final model.

Lastly, machine learning provides powerful methods for mapping numeric input to numeric output. However, not everything is reducible to numbers, especially not in healthcare. Primitive clinical characteristics, pictures, text, and images can all be expressed as 0's or 1's and thus easily be incorporated into a model, whereas human values pertinent to the patient remain irreducible to numbers. The model developed in **Chapter 5** predicts personalized estimates of expected survival with high accuracy and precision; however, it cannot grasp the personal and clinical implications associated with these predictions. As such, clinical decision-making can still be very different in two patients, even if the predicted outcomes are exactly the same. Clinicians should therefore be trained in considering the appropriate machine learning tools on case-by-case basis and interpreting the clinical implications associated with their predictions.

CONCLUSION

The thin line between treatment effectiveness and patient harms underpins the importance of tailoring clinical management to the individual brain tumor patient. Machine learning algorithms have the potential to unlock unique insights from large, complex data sources and effectively personalize clinical decision-making to the needs of the individual brain tumor patient. However, the automated nature comes at the cost of its interpretability, which can limit their clinical implementation and acceptance. Machine learning algorithms should be considered as an extension to statistical approaches and exist along a continuum determined by how much is specified by humans and how much is learnt by the machine. The choice of algorithm should be guided by the nature and complexity of the input data, as well as the desired level of human guidance and model interpretability. Although machine learning algorithms can produce highly accurate predictions based on high-dimensional data, clinicians and researchers should interpret the clinical implications of these predictions on case-by-case basis.

References

1. Ostrom QT, Gittleman H, Liao P, Vecchione-Koval T, Wolinsky Y, Kruchko C, et al. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. Neuro-Oncology. 2017 Nov 6;19(suppl_5):v1–88.
2. Bishop C. Pattern Recognition and Machine Learning [Internet]. New York: Springer-Verlag; 2006 [cited 2019 Mar 28]. (Information Science and Statistics). Available from: https://www.springer.com/in/book/9780387310732
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 27;521(7553):436–44.
4. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018 Apr 3;319(13):1317–8.
5. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians' working hours and lowers their career satisfaction. Int J Health Serv. 2014;44(4):635–42.
6. Rand V, Coleman C, Park R, Karar A, Khairat S. Towards Understanding the Impact of EHR-Related Information Overload on Provider Cognition. Stud Health Technol Inform. 2018;251:277–80.
7. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 04;380(14):1347–58.
8. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019 Jun 1;110:12–22.
9. Dabbish L, Stuart C, Tsay J, Herbsleb J. Social coding in GitHub: transparency and collaboration in an open software repository. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12 [Internet]. Seattle, Washington, USA: ACM Press; 2012 [cited 2019 Mar 13]. p. 1277. Available from: http://dl.acm.org/citation.cfm?doid=2145204.2145396