

# APPLIED MACHINE LEARNING *in* NEUROSURGICAL ONCOLOGY

Joeky T. Senders





# **Applied Machine Learning in Neurosurgical Oncology**

Joeky T. Senders

# Colophon

Cover design:	James Jardine   <a href="http://www.jamesjardine.nl">www.jamesjardine.nl</a>
Layout:	James Jardine   <a href="http://www.jamesjardine.nl">www.jamesjardine.nl</a>
Print:	Ridderprint   <a href="http://www.ridderprint.nl">www.ridderprint.nl</a>
ISBN:	978-94-6416-509-8

Work performed at Leiden University Medical Center, Leiden University, and Brigham and Women's Hospital, Harvard Medical School.

Email: [j.t.senders@gmail.com](mailto:j.t.senders@gmail.com)

The research fellowship was supported by grants from: Michaël Fonds, Fundatie van Renswoude, Sint Geertruidssleen, Nijbakker-Morra Stichting and kfHeijn Fonds.

Copyright © Joeky T. Senders, Amsterdam, The Netherlands. No part of this thesis may be reproduced, stored or transmitted in any form or by any means, without the prior permission of the author and the original copyright holder.

# **Applied Machine Learning in Neurosurgical Oncology**

PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Universiteit Leiden,  
op gezag van rector magnificus

prof. dr. ir. H. Bijl

volgens besluit van het college voor promoties te verdedigen op  
donderdag 27 januari 2022  
klokke 13:45 uur

door

Joeky Tamba Senders  
geboren te Utrecht

**Promotor**

prof. dr. W.C. Peul

**Copromotoren**

dr. mr. M.L.D. Broekman

dr. W.B. Gormley

BWH – Harvard Medical School

**Leden promotiecommissie**

prof. dr. M.J.N. Taphoorn

prof. dr. H. Putter

prof. dr. M.J.P. van Osch

prof. dr. P.A. Robe

UMCU – Universiteit Utrecht

prof. dr. A.M. May

UMCU – Universiteit Utrecht



# Table of Contents

Chapter 1	General introduction and thesis outline	9
-----------	---	---

## **Part I. Outcomes and risk factors in neurosurgical oncology**

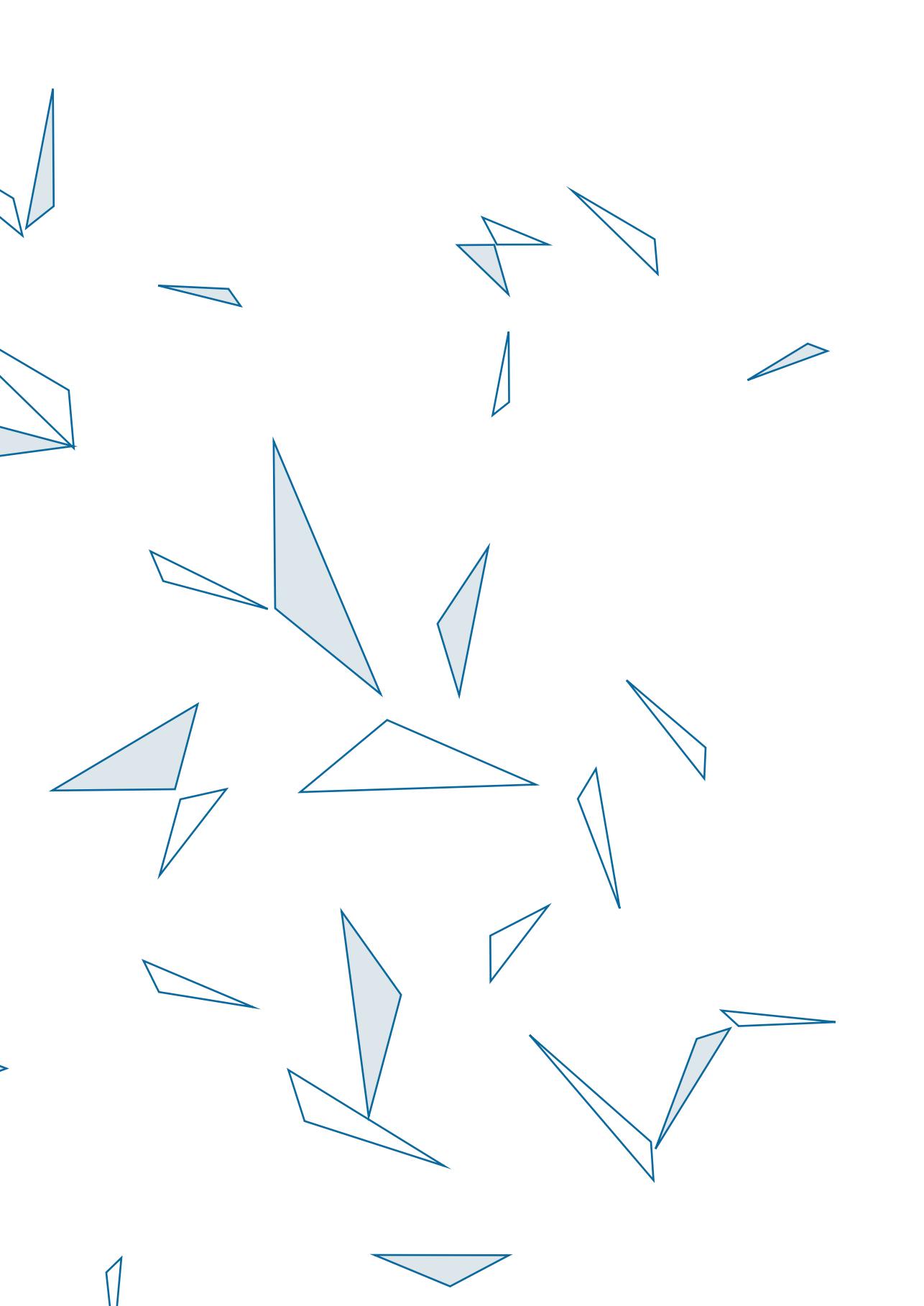
Chapter 2	Thirty-day outcomes after craniotomy for primary malignant brain tumors: A National Surgical Quality Improvement Program analysis <i>Neurosurgery 2018</i>	23
Chapter 3	Venous thromboembolism and intracranial hemorrhage after craniotomy for primary malignant brain tumors: A National Surgical Quality Improvement Program analysis <i>Journal of Neuro-Oncology 2018</i>	47
Chapter 4	Length of thromboprophylaxis in patients operated for a high-grade glioma: a retrospective study <i>World Neurosurgery 2018</i>	69

## **Part II. Predictive analytics in neurosurgical oncology**

Chapter 5	An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning <i>Neurosurgery 2020</i>	87
-----------	--	----

## **Part III. Natural language processing in neurosurgical oncology**

<b>Chapter 6</b>	Automating clinical chart review — an open-source natural language processing pipeline developed on free-text radiology reports of glioblastoma patients <i>JCO Clinical Cancer Informatics 2020</i>	<b>109</b>
<b>Chapter 7</b>	Natural language processing for automated quantification of brain metastases reported in free-text radiology reports <i>JCO Clinical Cancer Informatics 2019</i>	<b>127</b>
<b>Chapter 8</b>	Deep learning for natural language processing of free-text pathology reports — a comparison of learning curves <i>BMJ Innovations 2020</i>	<b>143</b>
<b>Chapter 9</b>	Summary	<b>161</b>
<b>Chapter 10</b>	General Discussion	<b>167</b>
<b>Appendices</b>	Nederlandse samenvatting	<b>183</b>
	Open-source code and software	<b>187</b>
	Authors and affiliations	<b>189</b>
	Acknowledgements	<b>191</b>
	Portfolio	<b>193</b>
	About the author	<b>201</b>



# 1

## General introduction and thesis outline



# GENERAL INTRODUCTION

## Neurosurgical oncology

### *Pathophysiology and epidemiology*

Malignant brain tumors are fast-growing neoplasms in the brain and can broadly be classified into primary and secondary (i.e., metastatic) brain tumors.<sup>1</sup> Primary brain tumors arise from within the brain, whereas secondary tumors originate elsewhere and spread to the brain as a result of hematogenous dissemination.<sup>2,3</sup> In the Netherlands, roughly 1,300 new patients are diagnosed with a primary malignant brain tumor each year.<sup>4,5</sup> Gliomas account for the majority (80-85%) and encompass a heterogenous group of brain tumors that originally evolve from astrocytes, oligodendrocytes, or ependymal cells.<sup>2</sup> These non-neural cell lines facilitate a wide range of supportive functions in the brain in addition to their structural support.<sup>6</sup> Due to the heterogenous nature and behavior of the disease, overall survival ranges between several months or years after diagnosis, yet all subtypes remain non-curative to date.<sup>7-9</sup> Glioblastoma constitutes the most common (~60%) and malignant glioma subtype with a median survival of 15 months after diagnosis despite improved surgical and adjuvant treatment strategies.<sup>10,11</sup> Brain metastases occur in approximately 6-17% of all cancer patients, and the incidence may be increasing as control of the systemic disease improves.<sup>3,12</sup> The most common primary tumors to metastasize to the brain are lung cancer, breast cancer, and melanoma accounting for 67-80% of all brain metastases.<sup>3</sup> The median survival in patients with brain metastases is typically in the order of months after diagnosis. Individual patient survival, however, varies widely depending on the age and functional status of the patient, the histological subtype and control of the primary tumor, and the number and intracerebral spread of brain metastases present.<sup>13</sup>

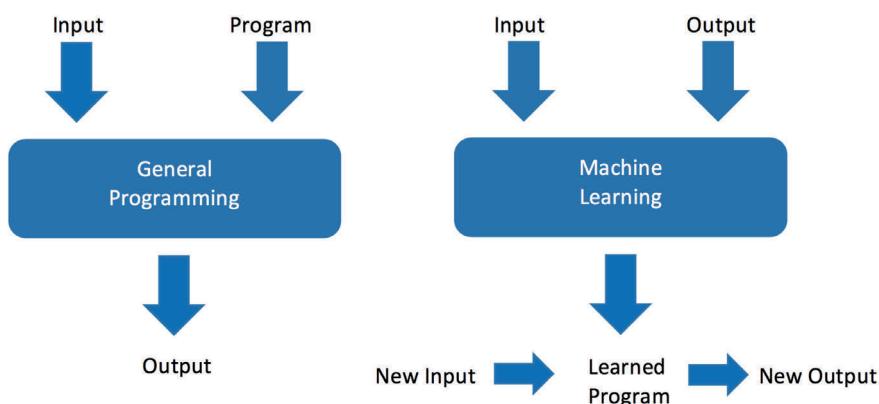
### *Clinical management*

The aggressive growth within healthy and functional brain tissue poses significant challenges with regards to the surgical and medical management of patients with a malignant brain tumor. Surgery is considered as the first line of treatment; however, the benefit of surgery should always be balanced against the risk of neurological deficits and mortality.<sup>7,14</sup> Adjuvant treatment largely depends on the histological and molecular subtype of the tumor. Histopathological examination of the resected tissue is therefore required to tailor clinical management to the needs of the individual patient. Standard of care for patients with a glioblastoma includes maximal safe resection followed by chemoradiation (i.e., radiotherapy with concomitant and adjuvant chemotherapy using temozolomide).<sup>7</sup> In the context of brain metastases, surgical resection should be considered in patients with a reasonable functional status and prognosis, a limited

number of large and superficial lesions, lesions with mass effect, and when control of the primary disease can be achieved.<sup>14</sup> Stereotactic radiosurgery or radiotherapy is used as adjuvant therapy, as well as monotherapy in patients with comorbidities precluding surgery or metastases that are irresectable due to their location and spread. Immunotherapy and hormone therapy can be viable treatment options as well, depending on the primary site and molecular subtype.<sup>14</sup>

## Applied machine learning

Clinical decision-making for patients with a malignant brain tumor should be performed on case-by-case basis given the heterogenous nature of the disease and the profound impact of surgical and adjuvant treatment strategies. As such, the clinical characteristics and personal preferences of the patient should be considered together with the biological profile of the tumor to tailor clinical management to the individual patient. The advent of the electronic health record system and complex diagnostic modalities used in neurosurgery provide a vast array of clinical information (e.g., free-text, imaging, histology, genomics) with the prospect of improving and personalizing future patient care. However, the high dimensionality and complexity of these data sources preclude clinicians from utilizing this information to its full potential.



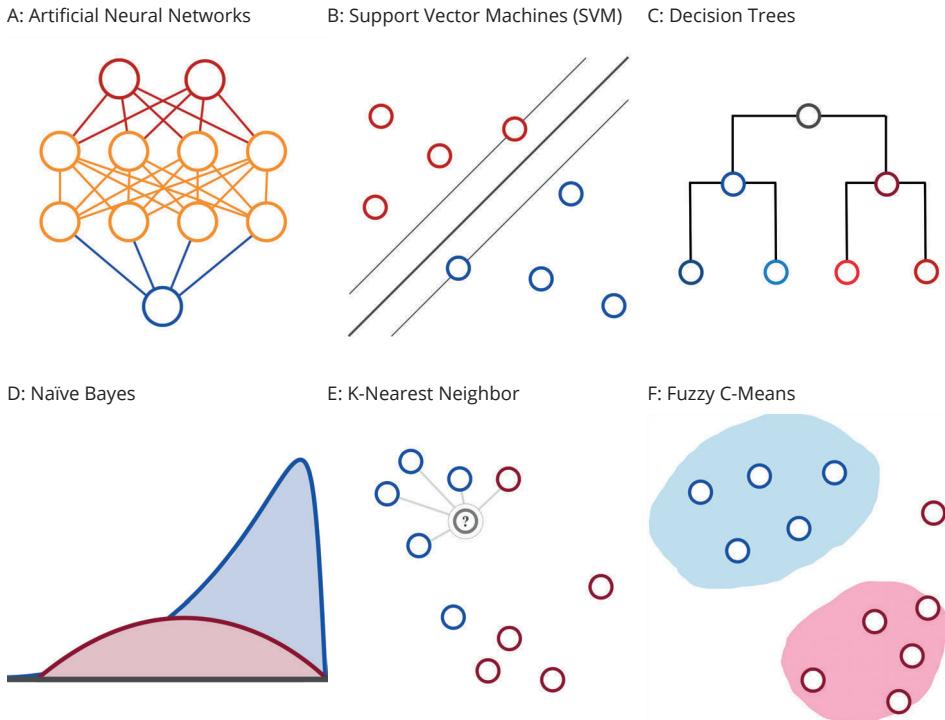
**FIGURE 1.** Difference between general programming and (supervised) machine learning.

Artificial intelligence is the branch of computer science concerned with the simulation of intelligent behavior in computers.<sup>15</sup> Machine learning is the branch of artificial intelligence that allows computer algorithms to learn from experience without explicitly being programmed.<sup>16</sup> As such, they have the ability to unlock unique insights

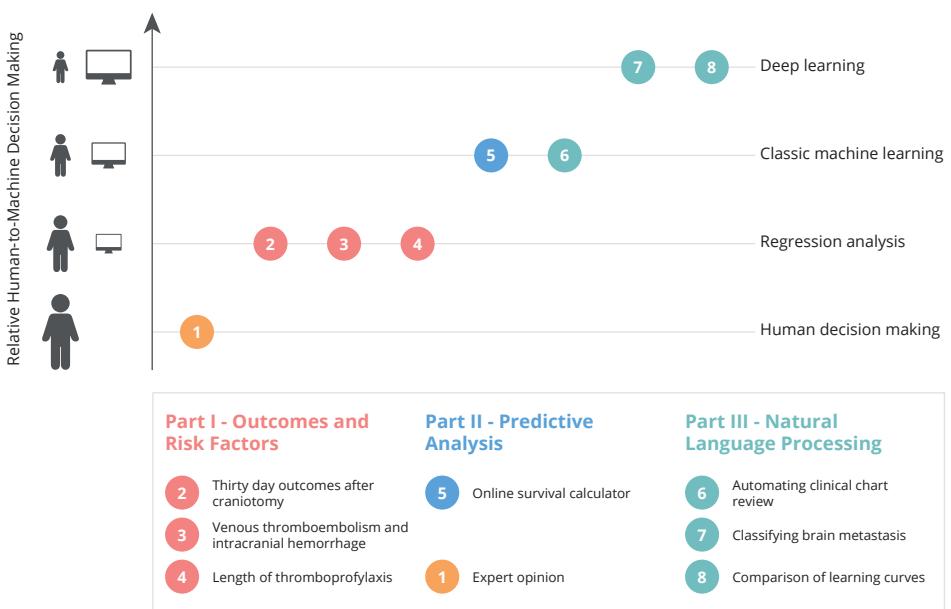
from complex data sources by learning patterns automatically, even those that are undetectable or meaningless for humans. Within the field of machine learning, a broad distinction can be made between supervised learning, unsupervised learning, and reinforcement learning.<sup>17</sup> Supervised learning algorithms learn from examples for which the desired output is known (i.e., labelled data) to develop a model that can compute predictions in new cases. Unsupervised learning techniques, on the other hand, seek to find similarities and patterns in unlabeled data. These algorithms can be valuable for identifying previously unknown clusters within the data. Reinforcement learning algorithms aim to determine the ideal behavior within a context or environment. It differs from supervised learning as it seeks to maximize the cumulative reward for series of actions instead of a single prediction.<sup>17</sup> The current thesis focusses on supervised algorithms and explores their utility in the clinical and scientific realms of neurosurgical oncology.

The ability to learn from known examples indicates the key difference between supervised machine learning and traditional programming. In traditional programming, a programmer manually writes a set of instructions – the program – to generate the desired output from a given set of inputs. In supervised machine learning, the input is provided together with the desired output, and computer algorithms are asked to derive the rules from the labeled training data. The product of this process is therefore not the desired output itself but a model that can predict the output in new observations (Figure 1). The automated learning process is an efficient way of analyzing large quantities of data, modelling hidden relationships in complex data sets, and adapting to changing environments. In the learning process, algorithms try to find the optimal combination of input variables (i.e., features) and weights given to these features in the model, thereby minimizing the difference between the predicted and observed outcomes. The mathematical structures of the most frequently used machine learning algorithms are briefly outlined in Figure 2.

Machine learning algorithms are founded upon statistical principles and should be considered as an extension of traditional statistical algorithms. They exist along a continuum determined by how much is specified by humans and how much is learnt by the machine, referred to as the machine learning spectrum.<sup>18</sup> For example, regression analysis on the low end of the machine learning spectrum requires more human guidance but provides valuable insights into the underlying predictive mechanisms, whereas deep learning on the high end of the spectrum can develop models from the raw data itself at the cost of model interpretability. The current thesis describes several studies along the continuum of the machine learning spectrum to derive knowledge from clinically-derived patient data and inform clinical decision-making in neurosurgical oncology (Figure 3).



**FIGURE 2.** Explanation of most frequently used prediction models. 1A: Artificial neural networks are inspired on the neural networks in the brain and organized in layers of interconnected nodes. The nodes in the upper layer (red) represent the input features, and the node in the lower layer (blue) represents the distinct output. The nodes in the 'hidden' (orange) and output layers (blue) base the value of their output on the total input they receive. The rapid increase in computing power in recent years has allowed researchers to develop artificial neural networks with many hidden layers and millions of parameters. This stacking of multiple layers, which is referred to as deep learning, allows the model to recognize complex patterns in higher-dimensional data. However, these models are also referred to as 'black box' algorithms as interpreting the predictive mechanisms can be challenging. 1B: Support vector machines classify data points by calculating the ideal straight line, the 'separating hyperplane'. Support vector machines select the hyperplane with the maximal distance to the nearest data point. A kernel function is mathematical trick that adds an extra dimension to the data. Non-separable 2-dimensional data, for example, could then be separated in a 3-dimensional space. 1C: Decision trees make predictions or classifications based on several input features with the use of bifurcating the feature space. These algorithms try to find the optimal features at which a split is made and the optimal value in case of a continuous feature. Random forests is an ensemble learning method that takes the mode of classes or the mean predictions of the individual trees, to avoid overfitting of a single tree. 1D: Naïve Bayes calculates the most likely outcome (blue) as a product of the a priori chance (red) and the conditional probabilities given by the individual features. Therefore, it assumes that the presence (or absence) of a feature is unrelated to the presence of any other feature, which is often not the case in real life. 1E: The K-Nearest Neighbors compares a data point with unknown class to its K nearest neighbors and determines its class as the most common class of its neighbors. For K=1, the algorithm assigns the class of a data point to the class of the single closest neighbor. 1F: Fuzzy C Means is an unsupervised learning algorithm that clusters data points based on their input features without having a desired output. The 'fuzzy' aspect gives the algorithm the flexibility to classify a data point to each cluster to a certain degree relating to the likelihood of belonging to that cluster.



**FIGURE 3.** The machine learning spectrum as it applies to the current thesis. Numbers 2 to 8 correspond to the chapters in the current thesis.

# THESIS OUTLINE

## Part I: Outcomes and risk factors in neurosurgical oncology

This part aims to provide an introduction into the characteristics and outcomes in patients undergoing brain tumor surgery. It comprises three retrospective studies that characterize the incidence and risk factors of postoperative morbidity and mortality.

**Chapter 2** examines the incidence and predictors of the occurrence of a major complication, extended length of stay, reoperation, readmission, and death within 30 days after surgery. Due to its high incidence and substantial impact on postoperative morbidity, a subsequent in-depth analysis (**Chapter 3**) was performed in the same cohort to characterize the rates, timing, and predictors of venous thromboembolism and intracranial hemorrhage. This chapter proposes a strategy for optimizing postoperative thromboprophylaxis – continuing prophylactic anticoagulation up to 21 days after surgery – which is evaluated on institutional data in **Chapter 4**.

## Part II: Predictive analytics in neurosurgical oncology

This part constitutes a study that evaluates the utility of statistical and machine learning algorithms for outcome prediction in neuro-oncology. **Chapter 5** provides a multimodal assessment of a variety of algorithms by evaluating their ability to predict survival in the individual glioblastoma patient based on structured demographic, socio-economic, clinical, and radiographic information. To facilitate the reproducibility and external validation of this chapter, the overall best performing model was deployed as an online survival calculator for glioblastoma patients.

## Part III: Natural language processing in neurosurgical oncology

This part evaluates the utility of natural language processing algorithms for automating clinical chart review in medical research. **Chapter 6** presents a natural language processing framework for automating the extraction of clinical information from free-text clinical reports. In this chapter, we analyze a text corpus of radiology reports of glioblastoma patients with a regression-based algorithm. Additionally, we characterize the association between model performance and statistical properties of the variables of interest to provide insight into the methodological boundaries and variables eligible for clinical text mining. In **Chapter 7**, we develop several models to classify radiology reports of brain metastases patients into reports that describe solitary versus multiple metastases. This chapter includes an extensive comparison between various statistical, classical machine learning, and deep learning algorithms to provide insight into their

relative performance and utility for medical text analysis. **Chapter 8** compares the learning curves of various algorithms in determining the histopathological diagnosis of brain tumor patients based on free-text pathology reports. These learning curves elucidate the learning efficiency of various algorithms, as well as the yield of natural language processing in clinical research. Furthermore, a modified deep learning architecture is developed in this chapter and its performance compared to conventional deep learning architectures and regression-based algorithms.

## **Summary and general discussion**

**Chapter 9** synthesizes the most important findings of the current thesis. **Chapter 10** addresses the implications of these finding with regards to the clinical care of neurosurgical and neuro-oncological patients. It provides recommendations for future machine learning research in healthcare. All custom computer code and software developed throughout this thesis have been made publicly-available and their associated URL-links can be found in the **Open-source code and software appendix**.

## References

1. Ostrom QT, Cioffi G, Gittleman H, Patil N, Waite K, Kruchko C, et al. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012–2016. *Neuro-Oncology*. 2019 Nov 1;21(Supplement\_5):v1–100.
2. Schwartzbaum JA, Fisher JL, Aldape KD, Wrensch M. Epidemiology and molecular pathology of glioma. *Nature Clinical Practice Neurology*. 2006 Sep;2(9):494–503.
3. Nayak L, Lee EQ, Wen PY. Epidemiology of Brain Metastases. *Curr Oncol Rep*. 2012 Feb 1;14(1):48–54.
4. neuro-oncologie [Internet]. [cited 2019 Jul 8]. Available from: <https://www.iknl.nl/oncologische-zorg/tumorteams/neuro-oncologie>
5. Nederlandse Kankerregistratie [Internet]. [cited 2019 Jul 5]. Available from: <https://www.cijfersoverkanker.nl/>
6. Ostrom QT, Gittleman H, Truitt G, Boscia A, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011–2015. *Neuro Oncol*. 2018 Oct 1;20(suppl\_4):iv1–86.
7. Weller M, van den Bent M, Hopkins K, Tonn JC, Stupp R, Falini A, et al. EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *The Lancet Oncology*. 2014 Aug 1;15(9):e395–403.
8. Weller M, van den Bent M, Tonn JC, Stupp R, Preusser M, Cohen-Jonathan-Moyal E, et al. European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas. *The Lancet Oncology*. 2017 Jun 1;18(6):e315–29.
9. Gorlia T, Bent MJ van den, Hegi ME, Mirimanoff RO, Weller M, Cairncross JG, et al. Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981–22981/CE.3. *The Lancet Oncology*. 2008 Jan 1;9(1):29–38.
10. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJB, et al. Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine*. 2005 Mar 10;352(10):987–96.
11. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The Lancet Oncology*. 2009 May 1;10(5):459–66.
12. Davis FG, Dolecek TA, McCarthy BJ, Villano JL. Toward determining the lifetime occurrence of metastatic brain tumors estimated from 2007 United States cancer incidence data. *Neuro Oncol*. 2012 Sep 1;14(9):1171–7.
13. Stelzer KJ. Epidemiology and prognosis of brain metastases. *Surg Neurol Int*. 2013 May 2;4(Suppl 4):S192–202.
14. Soffietti R, Abacioglu U, Baumert B, Combs SE, Kinhult S, Kros JM, et al. Diagnosis and treatment of brain metastases from solid tumors: guidelines from the European Association of Neuro-Oncology (EANO). *Neuro Oncol*. 2017 Feb 1;19(2):162–74.
15. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015 May 28;521(7553):452–9.
16. 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>
17. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. 2015 Jul 17;349(6245):255–60.
18. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018 Apr 3;319(13):1317–8.

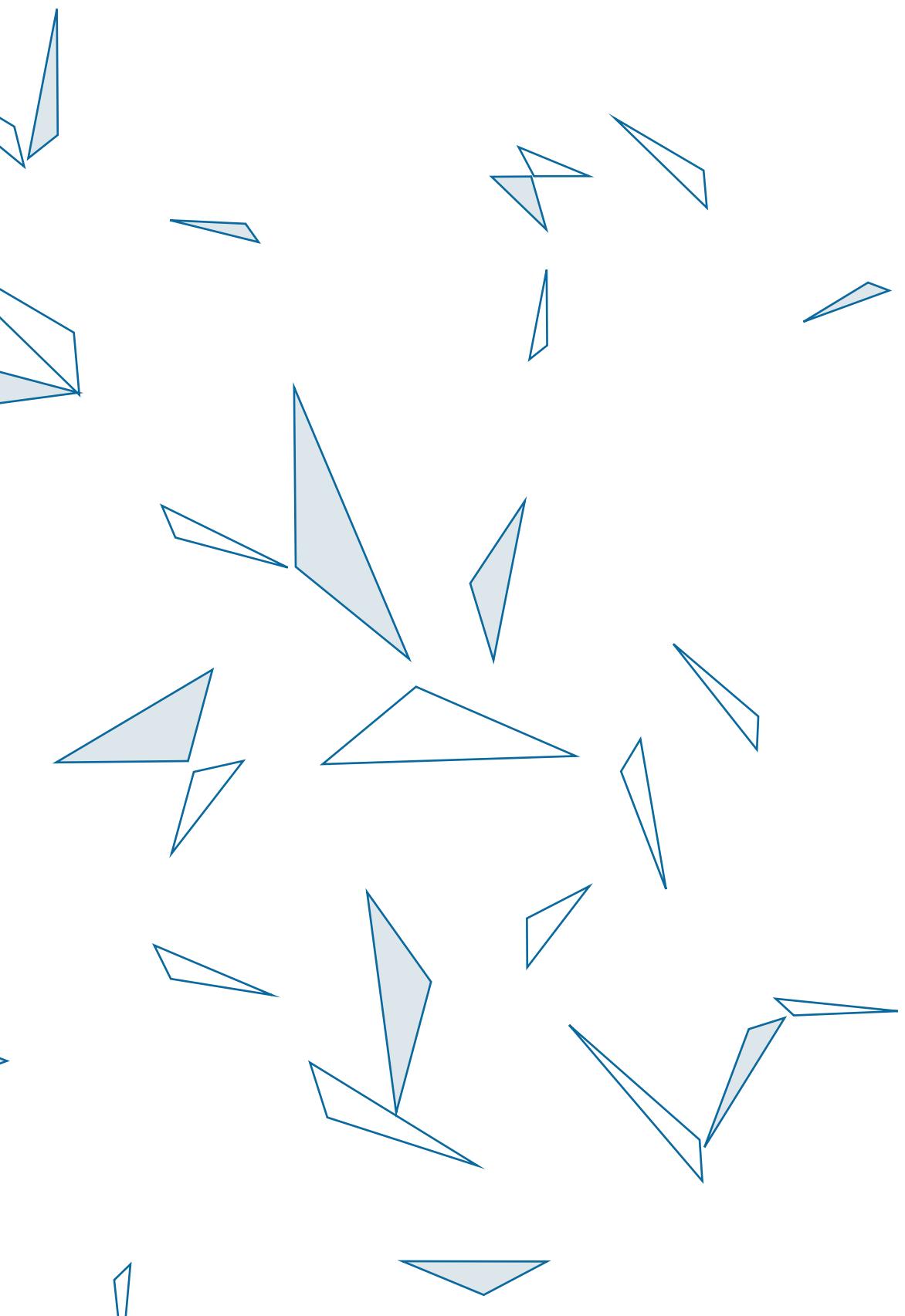




# PART I

---

**Outcomes and risk factors in  
neurosurgical oncology**



# 2

## Thirty-day outcomes after craniotomy for primary malignant brain tumors: a National Surgical Quality Improvement Program analysis

---

Joeky T. Senders, Ivo S. Muskens, David J. Cote, Nicole H. Goldhaber, Hassan Y. Dawood, William B. Gormley, Marike L.D. Broekman, Timothy R. Smith

NEUROSURGERY. 2018 DEC 1;83(6):1249-1259

# **Abstract**

## **Introduction**

Despite improved perioperative management, the rate of postoperative morbidity and mortality after brain tumor resection remains considerably high. This study assesses the rates, causes, timing, and predictors of major complication, extended length of stay ( $>10$  days), reoperation, readmission, and death within 30 days after craniotomy for a primary malignant brain tumor.

## **Methods**

Patients were extracted from the National Surgical Quality Improvement Program registry (2005-2015) and analyzed using multivariable logistic regression.

## **Results**

7376 patients were identified, of which 948 (12.9%) experienced a major complication. The most common major complications were reoperation (5.1%), venous thromboembolism (3.5%), and death (2.6%). Furthermore, 15.6% stayed longer than ten days, and 11.5% were readmitted within 30 days after surgery. The most common reasons for reoperation and readmission were intracranial hemorrhage (18.5%) and wound-related complications (11.9%), respectively. Multivariable analysis identified older age, higher body mass index (BMI), higher American Society of Anesthesiologists (ASA)-classification, dependent functional status, elevated preoperative white blood cell count (WBC,  $>12,000$  cells/mm $^3$ ), and longer operative time as predictors of major complication (all  $p<0.001$ ). Higher ASA-classification, dependent functional status, elevated WBC, and ventilator dependence were predictors of extended length of stay (all  $p<0.001$ ). Higher ASA-classification and elevated WBC were predictors of reoperation (both  $p<0.001$ ). Higher ASA-classification and dependent functional status were predictors of readmission (both  $p<0.001$ ). Older age, higher ASA-classification, and dependent functional status were predictors of death (all  $p<0.001$ ).

## **Conclusion**

This study provides a descriptive analysis and identifies predictors for short-term complications, including death, after craniotomy for a primary malignant brain tumor.

## Introduction

Despite improvements in the treatment of primary malignant brain tumors, these invasive cancers continue to result in significant morbidity and mortality. The estimated median survival is as short as 15 months for patients with a glioblastoma, the most common and deadliest type of primary malignant brain tumor.<sup>1</sup>

Currently, the standard of care for newly diagnosed primary malignant brain tumors is maximal safe resection of the tumor, typically followed by adjuvant chemotherapy and radiation.<sup>2</sup> Balancing between maximal resection and preservation of neurological functioning can be challenging due to the infiltrative nature of these tumors. Additionally, the short-term postoperative course is frequently complicated by major adverse events, often resulting in extended length of stay, reoperation, and readmission.<sup>3</sup>

Identifying the most predominant patient demographics and procedural related factors associated with postoperative morbidity and mortality would be beneficial for improving patient selection and tailoring postoperative management to a patient's individual risk profile. Although many multicenter studies have reported on short-term outcomes after craniotomy for brain tumors,<sup>4-14</sup> only few have focused on patients with a primary malignant brain tumor.<sup>10-12,15</sup> These studies provide valuable but limited insight into the direct postoperative course because they often assess a specific subset of outcome measures only and use varying inclusion criteria.

The current study aims to provide an overall picture of the postoperative course of this patient population by assessing the rates, causes, timing, and predictors of major short-term outcomes after craniotomy for a primary malignant brain tumor. For this purpose, the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) database was reviewed to investigate the occurrence of major complications, extended length of stay, reoperation, readmission, and death within the first 30 days following craniotomy. The results of this multicenter study can aid to optimize postoperative management of patients undergoing surgical resection of a primary malignant brain tumor.

## Methods

### Data source

All patients who underwent a craniotomy for a primary malignant brain tumor were extracted from the NSQIP registry (2005-2015). This registry tracks patients prospectively

for 30 days after their surgery and includes data from over 600 hospitals across the United States.<sup>16</sup> This validated dataset is collected by trained surgical reviewers at each site using a standardized protocol and includes the most common postoperative complications, length of stay, occurrence of reoperations and readmissions together with associated reasons, and mortality.<sup>17</sup> The NSQIP registry has previously been used to study surgical outcomes after neurosurgical procedures.<sup>6,8,17-26</sup> Our institutional review board has exempted this study from review. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines were used in this study.

## **Inclusion criteria**

Patients were included who met the following criteria: 1) aged 18 years or older; 2) Current Procedural Terminology (CPT) code(s) indicating craniotomy for surgical resection of brain tumors (CPT: 61500, 61510, 61512, 61518, 61519, 61520, 61521, 61526, and 61530); 3) postoperative diagnosis indicative of primary malignant brain tumor according to International Classification of Diseases, Ninth Revision [ICD-9]: 191.x. CPT codes are medical codes maintained by the American Medical Association to communicate uniform information and billing on medical, surgical, and diagnostic procedures. Similarly, ICD-9 codes are also used for uniform communication; however, ICD-9 codes refer to diagnoses, whereas CPT codes refer to medical services.

## **Covariates**

Covariates were extracted and analyzed as follows. Age, body mass index (BMI), and operative time were examined continuously. American Society of Anesthesiologists (ASA) classification was examined categorically (I-II, III, IV-V). Other evaluated preoperative patient characteristics included: gender, race, smoking status during the past year, dyspnea, chronic obstructive pulmonary disease (COPD), hypertension requiring medication, bleeding disorders, diabetes mellitus (insulin dependent vs. non-insulin dependent), steroid usage, recent congestive heart failure, preoperative functional status (dependent vs. independent), and systemic sepsis. Preoperative laboratory values were also extracted and categorized according to standard reference ranges and clinical significance. When 'elevated' or 'decreased' laboratory groups consisted of a very low number of cases or when its association with the independent variables did not significantly differ from the normal group, this category was deemed to be clinically non-significant and merged with the 'normal' group. This reduces the degrees of freedom and preserves statistical power, resulting in the following categories: creatinine (<1.4 mg/dL vs. ≥1.4 mg/dL), hematocrit (<36% vs. ≥36%), platelet count (100,000-450,000/µL, <100,000/µL, or >450,000/µL), sodium (135-145 mEq/L, <135 mEq/L, or >145 mEq/L), and white blood cell count (WBC, ≤12,000/µL or >12,000/

$\mu\text{L}$ ). Covariates present in less than one percent of cases or with more than ten percent missing data were excluded from the analysis. Cases with missing data in one of the variables of the multivariable analysis were excluded from the analysis.

## Outcomes

Major complications, extended length of stay, reoperations and readmission with related reasons, and mortality within 30-days were extracted as primary endpoints. Based on a previously published definition, major complication was defined as cardiovascular accident, cardiac arrest, myocardial infarction, deep venous thrombosis, pulmonary embolism, unplanned intubation, failure to wean from ventilator, acute renal failure, sepsis, septic shock, deep incisional or organ space surgical site infection, return to the operating room, or death.<sup>27</sup> Extended length of stay was defined as a total hospital stay of more than ten days. Unplanned reoperation data has been collected by NSQIP since its inception in 2005, but unplanned readmission data has been collected since 2011 regardless of the hospital to which the patient was readmitted. Associated procedures for reoperation, based on CPT codes, and reasons for readmission, based on ICD-9 codes, have been collected since 2012. Non-routine discharge was defined as discharge to another acute care facility, skilled nursing facility, or rehabilitation center.

## Statistical analysis

Statistical analyses were conducted using R 3.3.3 (R Core Team, Auckland, New Zealand). Descriptive statistics were performed on baseline demographics after categorization, together with univariable analysis for each of the primary endpoints by means of logistic regression. Multivariable logistic regression models were constructed using variables screened by univariable analysis for occurrence of major complication, extended length of stay, reoperation, readmission, and death. All potential predictors ( $p<0.05$  in the univariable analysis) were included in the multivariable analysis, but age and gender were included automatically. Potential predictors were excluded from the final model if they demonstrated multicollinearity or had a relative low contribution to model fit. Due to the relatively large sample size of the study population, a  $p<0.0015$  was considered as significant, to decrease the chance of Type I (false-positive) error. This critical value was based on a Bonferroni correction for multiple testing with 33 degrees of freedom ( $0.05/33=0.0015$ ). A confirmatory analysis was performed for every multivariable model, in which missingness was coded as an additional group to verify if the missing group significantly affected the results. The  $\beta$ -coefficients of the continuous variables in the final model were scaled to represent the odds ratios and confidence intervals of meaningful and interpretable units for age (per ten-year increase), BMI (per five  $\text{kg}/\text{m}^2$  increase), and operative time (per 60-minute increase).

# Results

## Demographics of study population

7376 NSQIP-reported patients underwent craniotomy for resection of a primary malignant brain tumor during the study period. Comorbidities, demographics, and preoperative laboratory values separated by occurrence of a major complication are summarized in Table 1.

## Outcomes

16.4% of patients experienced any complication, and 12.9% of patients experienced a major complication in the first 30 days after surgery (Table 2). The most common major complications were reoperation (5.1%), venous thromboembolism (3.5%), and death (2.6%). Most major complications occurred within the first two weeks after surgery (median time to major complication: nine days), and 82.3% occurred during hospitalization. The median hospital stay was four days (interquartile range three to eight days), and 15.6% of patients stayed longer than ten days. During the initial 30-days after surgery, 5.1% of patients required reoperation at a median of 12 days after surgery. The most common reasons for reoperation were intracranial hemorrhage (ICH, 18.5%), hydrocephalus (17.8%), and resection of residual tumor tissue (16.4%).

Readmission occurred in 11.5% of cases, at a median of 12 days after discharge. The most common reasons for readmission were wound-related complications (11.9%), seizures (8.8%), and venous thromboembolism (7.4%) (Figure 1). Death within 30 days after surgery occurred in 2.6% of cases, of which 37.9% occurred during the initial hospital stay. The incidence of major complications, extended length of stay, reoperation, and death remained fairly consistent from 2009 to 2015 (Figure 2). After an initial drop in 2012, the readmission rate also remains fairly consistent; however, no readmission data was available for patients operated on before 2011.

**TABLE 1.** Demographics and preoperative comorbidities of NSQIP patients undergoing craniotomy for a primary malignant brain tumor, compared by the occurrence of a major complication.

Characteristic (%)	Definition	Total percentage (n=7376)	Major complication (n=948)	No complication (n=6428)	OR	95% CI	p
Mean age <sup>a</sup>	Years ± SD	54.5±15.6	57.3±15.5	54.1±15.6	1.15 <sup>d</sup>	1.10-1.20	<.001
Sex <sup>a</sup>					Ref.	-	-
Female	42.3	41.5	42.5				
Male	57.7	58.5	57.5	1.04	0.91-1.20	.56	
Race <sup>c</sup>					Ref.	-	-
White	91.8	91.3	91.9				
Black	4.7	5.6	4.6	1.22	0.86-1.69	.25	
Asian	2.9	2.7	3.0	0.90	0.54-1.41	.66	
Other	0.5	0.4	0.5	0.83	0.20-2.36	.75	
ASA-classification <sup>b</sup>					Ref.	-	-
I-II	27.7	16.6	29.4				
III	59.2	62.8	58.7	1.90	1.58-2.28	<.001	
IV-V	13.0	20.6	11.9	3.05	2.43-3.84	<.001	
Mean BMI <sup>b</sup>	kg/m <sup>2</sup> ± SD	28.4±6.2	29.2±6.5	28.3±6.1	1.26 <sup>e</sup>	1.13-1.40	<.001
History of CHF		0.2 <sup>d</sup>	0.3	0.2	1.36	0.39-4.70	.63
Hypertension		35.4	44.4	34.1	1.55	1.35-1.77	<.001
Smoking		17.1	15.7	17.3	0.89	0.74-1.07	.21
Emergency case		6.5	10.2	6.0	1.78	1.40-2.24	<.001
Admitted not from home <sup>b</sup>		18.1	25.1	17.2	1.61	1.37-1.89	<.001
History of COPD		2.4	3.8	2.1	1.80	1.24-2.61	.002
Ventilator dependent		1.1	4.2	0.7	6.39	4.14-9.86	<.001
Dialysis		0.1 <sup>d</sup>	0.3	0.1	2.91	0.75-11.3	.11
Renal failure		0.1 <sup>d</sup>	0.3	0.1	5.10	1.14-22.8	.02
Weight loss		1.7	3.1	1.5	2.12	1.39-3.24	.001

TABLE 1. Continued.

Characteristic (%)	Definition	Total percentage (n=7376)	Major complication (n=948)	No complication (n=6428)	OR	95% CI	p
Transfusion		0.2 <sup>d</sup>	0.5	0.2	3.09	1.07-8.92	.03
Bleeding disorder		2.2	4.3	1.9	2.37	1.66-3.41	<.001
Dyspnea		2.6	4.2	2.3	1.86	1.30-2.65	.001
Insulin dependent DM		3.9	6.4	3.5	1.91	1.42-2.54	<.001
Pre-op steroid usage		16.6	20.9	16.0	1.38	1.17-1.64	<.001
Dependent functional status <sup>a</sup>		5.1	10.1	4.4	2.41	1.88-3.07	<.001
Preoperative sodium <sup>b</sup>	135-145 mEq/L	89.9	86.7	90.4	Ref.	-	-
	<135 mEq/L	9.1	11.6	8.7	1.38	1.10-1.72	.004
	>145 mEq/L	1.0	1.7	0.9	1.94	1.07-3.31	.02
Preoperative creatinine <sup>b</sup>	≥1.4 mg/dL	4.4	6.2	4.3	1.45	1.08-1.95	.01
Preoperative WBC <sup>c</sup>	>12000/ $\mu$ L	24.8	33.3	23.5	1.63	1.40-1.89	<.001
Preoperative hematocrit <sup>b</sup>	<36%	11.9	14.3	11.1	1.31	1.07-1.60	.007
Platelets <sup>b</sup>	100-450	97.5	96.1	97.7	Ref.	-	-
	<100	1.3	2.4	1.1	2.12	1.28-3.39	.002
	>450	1.2	1.5	1.1	1.33	0.72-2.30	.33
Median operative time <sup>a</sup>	Minutes (IQR)	179(123-250)	187(130-258)	178(122-197)	1.19 <sup>f</sup>	1.03-1.37	<.001
No general anesthesia <sup>a</sup>		5.9	5.2	6.0	0.85	0.62-1.14	.30

Abbreviations: ASA=American Society of Anesthesiologists, BMI=Body Mass Index; CHF=Chronic Heart Failure; CI=Confidence Interval; DM=diabetes mellitus; IQR=interquartile range; OR=odds ratio; p=p-value;  $\mu$ L=Micro-Liter; SD=Standard Deviation; WBC=white blood cell count

For all tests,  $p<.0015$  was considered significant.

<sup>a</sup> Independent variable with <1% missing data.

<sup>b</sup> Independent variable with 1-10% missing data.

<sup>c</sup> Independent variable with >10% missing data.

<sup>d</sup> Inflated  $\beta$ -coefficients to odds ratio per 10 years increase.

<sup>e</sup> Inflated  $\beta$ -coefficients to odds ratio per 5 kg/m<sup>2</sup> increase.

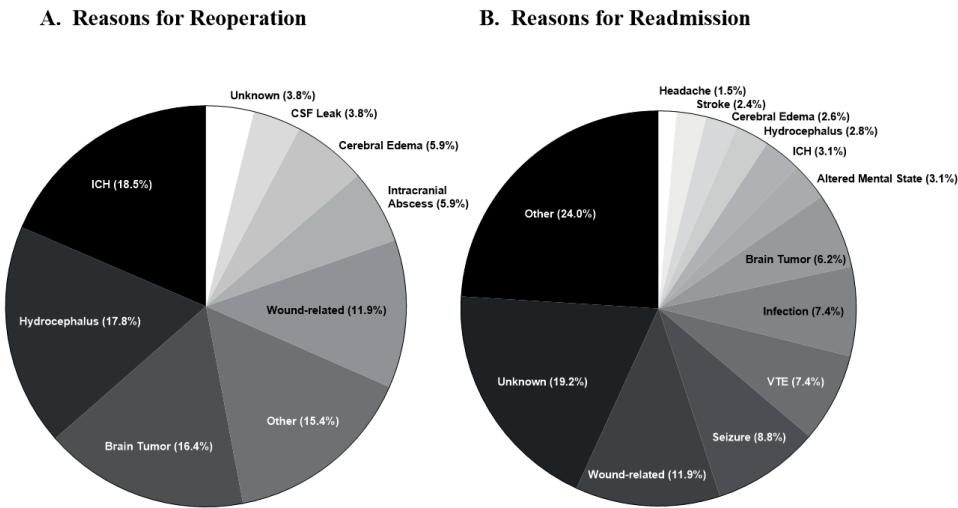
<sup>f</sup> Inflated  $\beta$ -coefficients to odds ratio per 60 minutes increase

**TABLE 2.** Complication incidence rates and time-to-event data within 30 days after craniotomy for primary malignant brain tumors.

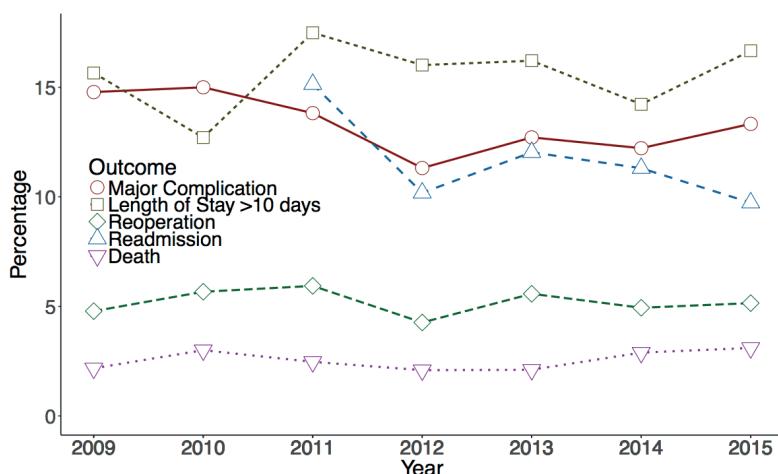
Complication	Thirty-day cumulative incidence (%)	Median time-to-event in days (IQR)	Occurrence during initial hospitalization (%)
Major complication <sup>a</sup>	12.9	9.0 (2.0-18.0)	82.3
Any complication	16.4	9.0 (2.0-18.0)	76.3
Length of stay	-	4.0 (3.0-8.0)	-
Length of stay > 10 days	15.6	-	-
Death within 30 Days	2.6	16.5 (10.0-23.3)	37.9
Reoperation	5.1	12.0 (3.0-20.3)	51.6
Readmission	11.5	16.0 (10.0-23.0)	-
Non-routine hospital discharge	31.1	5 (3.0-9.0)	-
Neurological complications			
Cardiovascular accident	1.4	2.0 (1.0-8.5)	73.3
Cardiovascular complications	0.5	9.5 (2.0-17.5)	63.2
Cardiac arrest	0.3	10.5 (3.25-17.0)	66.7
Myocardial infarction	0.2	6.5 (1.25-18.0)	56.3
Hematologic complications	3.7	14.5 (8.0-22.0)	33.5
Deep venous thrombosis	2.6	13.0 (7.0-21.0)	39.6
Blood transfusion	0.2	NA	NA
Pulmonary embolism	1.5	17.0 (9.0-23.0)	27.1
Pulmonary complications	3.0	3.0 (2.0-9.0)	78.2
Unplanned intubation	2.1	5.0 (1.0-12.5)	71.2
Failure to wean from ventilator	1.9	3.0 (2.0-6.0)	91.5
Renal complications			
Renal failure	0.1	12.0 (10.0-15.0)	40.0
Infectious complications	6.3	11.0 (6.0-18.0)	43.1
Surgical site infection	2.2	15.5 (11.0-21.0)	20.1
Pneumonia	1.6	8.0 (4.0-16.5)	65.3
Urine tract infection	2.3	10.0 (7.0-15.0)	48.5
Sepsis/Septic Shock	2.1	11.0 (7.0-18.0)	46.8

Abbreviations: IQR=interquartile range, NA=not available.

<sup>a</sup> Defined as cardiovascular accident, cardiac arrest, myocardial infarction, deep venous thrombosis, pulmonary embolism, unplanned intubation, failure to wean from ventilator, acute renal failure, sepsis, septic shock, deep incisional or organ space surgical site infection, return to the operating room, or death within 30 days.



**FIGURE 1.** Reasons for A) reoperation and B) readmission among NSQIP-reported patients undergoing resection of a primary malignant brain tumor. Abbreviations: CSF=cerebrospinal fluid; ICH=intracranial hemorrhage; VTE=venous thromboembolism.



**FIGURE 2.** Complication rates per year. Incidence rates for 2005-2008 were not included in this figure due to the low number of patients ( $n < 100$  per year). Readmission data was collected since 2011.

Multivariable analysis of primary endpoints: Multivariable analysis identified older age, higher BMI, higher ASA-classification, dependent functional status, elevated WBC, and longer operative times as independent predictors of major complication

(all  $p<0.001$ , Table 3). Older age, higher ASA-classification, dependent functional status, elevated WBC, low hematocrit, and ventilator dependence were predictors of hospital stay beyond ten days (all  $p<0.001$ ). Higher ASA-classification and elevated WBC were predictors of reoperation within 30 days (both  $p<0.001$ ) (Table 4). Higher ASA-classification and dependent functional status were predictors of readmission (both  $p<0.001$ ). Older age, higher ASA-classification, and dependent functional status were predictors of 30-day mortality (all  $p<0.001$ ) (Table 5). Missingness ranged between 6.2% and 10.6% depending on the specific outcome measure studied. Including missingness as an additional group did not significantly alter the results.

## Discussion

Despite improved perioperative management, the overall rate of short-term postoperative major complications, extended length of stay, reoperation, readmission, and mortality for patients treated with a craniotomy for primary malignant brain tumors remains considerably high. This multicenter registry study provides a descriptive analysis and identifies predictors of 30-day postoperative outcomes after craniotomy for these lesions.

The patient population in this study was technically classified as all those with primary malignant brain tumors based on ICD-9 codes. Gliomas represent close to 80% of primary malignant brain tumors;<sup>1</sup> however, there is no standard ICD-9 code specific for glioma. The Central Brain Tumor Registry of United States (CBTRUS) argues that multiple combinations of ICD histology codes can be used to define gliomas, and their approach was modeled in this study.<sup>1</sup> Therefore, these results are primarily applicable to glioma patients and should be put in the context of previous outcome research on glioma patients.

Many multicenter studies have reported on short-term outcomes after craniotomy for brain tumors;<sup>4-14</sup> however, few have specifically analyzed primary malignant brain tumor or glioma patients.<sup>10-12,14,15</sup> Previous multicenter studies found similar length of stay, readmission rate, and mortality rate in glioma patients<sup>10-12,15</sup> and brain tumor patients overall, including primary benign brain tumors, metastases, meningiomas, and cranial nerve tumors.<sup>7,9,13,28</sup> The rates of overall and major complications in the current study were similar to other studies analyzing glioma patients,<sup>11,14</sup> but fall in the upper range of studies that analyzed all brain tumor patients.<sup>7,13</sup> This suggests that the postoperative course of glioma patients is more frequently complicated by adverse events compared

**TABLE 3.** Multivariable logistic regression identifying predictors of a major complication and extended length of stay

Predictor	Definition	Major Complication			Hospital Stay >10 days		
		OR	95% CI	p	OR	95% CI	p
Age	Per 10 years increase	1.09 <sup>a</sup>	1.04-1.15	<.001	1.12 <sup>a</sup>	1.07-1.17	<.001
Gender	Female	Ref.	-	-	Ref.	-	-
BMI	Male	1.05	0.90-1.22	.52	0.96	0.84-1.10	.56
ASA-classification	Per 5 Kg/m <sup>2</sup> increase	1.11 <sup>b</sup>	1.05-1.17	<.001	-	-	-
	I-II	Ref.	-	-	Ref.	-	-
	III	1.59	1.30-1.94	<.001	2.09	1.73-2.55	<.001
	IV-V	2.42	1.90-3.13	<.001	3.46	2.74-4.39	<.001
Functional status	Independent	Ref.	-	-	Ref.	-	-
	Dependent	2.09	1.60-2.71	<.001	2.48	1.96-3.13	<.001
WBC count	≤12,000/ $\mu$ L	Ref.	-	-	Ref.	-	-
	>12,000/ $\mu$ L	1.52	1.29-1.78	<.001	1.59	1.38-1.84	<.001
Hematocrit	≥36%	Ref.	-	-	Ref.	-	-
	<36%	-	-	-	1.53	1.26-1.84	<.001
Ventilator dependence	No	Ref.	-	-	Ref.	-	-
	Yes	-	-	-	3.77	2.23-6.40	<.001
Operative time	Per 60 minutes increase	1.13 <sup>c</sup>	1.09-1.17	<.001	-	-	-
Model fit	AUC	0.64	0.62-0.66	<.001	0.68	0.66-0.70	<.001

ASA=American Society of Anesthesiologists, AUC=Area under the curve; BMI=Body Mass Index, CI=confidence interval; p=p-value; OR=odds ratio; WBC=white blood cell; For all tests, p<.0015 was considered significant.

<sup>a</sup>Inflated β-coefficients to odds ratio per 10 years increase.

<sup>b</sup>Inflated β-coefficients to odds ratio per 5 kg/m<sup>2</sup> increase.

<sup>c</sup>Inflated β-coefficients to odds ratio per 60 minutes increase.

**TABLE 4.** Multivariable logistic regression identifying predictors of reoperation and readmission.

Predictor	Definition	Reoperation			Readmission		
		OR	95% CI	p	OR	95% CI	p
Age	Per 10 years increase	0.92 <sup>a</sup>	0.86-0.99	.03	1.03 <sup>a</sup>	0.98-1.09	.22
Gender	Female	Ref.	-	-	Ref.	-	-
	Male	1.10	0.88-1.38	.38	0.97	0.83-1.14	.76
BMI	Per 5 Kg/m <sup>2</sup> increase	1.09 <sup>b</sup>	1.01-1.19	.03	1.05 <sup>b</sup>	0.99-1.11	.13
ASA-classification	I-II	Ref.	-	-	Ref.	-	-
	III	1.87	1.38-2.54	<.001	1.45	1.19-1.77	<.001
	IV-V	2.91	1.99-4.24	<.001	1.12	0.83-1.51	.43
Functional status	Independent	-	-	-	Ref.	-	-
	Dependent	-	-	-	2.04	1.48-2.76	<.001
WBC count	≤12,000/ $\mu$ L	Ref.	-	-	-	-	-
	>12,000/ $\mu$ L	1.54	1.22-1.94	<.001	-	-	-
Model fit	AUC	0.62	0.59-0.65	<.001	0.57	0.55-0.59	<.001

ASA=American Society of Anesthesiologists, AUC=Area under the curve; BMI=Body Mass Index, CI=confidence interval; OR=odds ratio; p=p-value; WBC=white blood cell; For all tests, p<.001 was considered significant.

<sup>a</sup> Inflated β-coefficients to odds ratio per 10 years increase.

<sup>b</sup> Inflated β-coefficients to odds ratio per 5 kg/m<sup>2</sup> increase.

to other types of brain tumors. Additionally, the reoperation rate in glioma patients has not been described in previous multicenter studies. Multicenter studies including brain tumor patients in general demonstrated similar rates of reoperation.<sup>6,13,29</sup>

Single-center or single-surgeon studies have reported a wide range of complication rates (4.3-14%) and mortality rates (0-3.7%) after craniotomy for resection of primary malignant brain tumors.<sup>15,30-33</sup> Despite the 30-day timeframe used, the complication and mortality rates determined in this multicenter study (12.9% and 2.6% respectively) fall into the upper ends of the results from observational studies that used a longer follow up. Although these single institutional observational studies provide valuable insight, they can be less generalizable due to their retrospective nature, small sample sizes, and single surgeon or single institutional experience. Because observational studies are often from academic tertiary centers with more specific expertise, the actual rates of postoperative morbidity and mortality may be underestimated in these observational studies. Two studies have previously demonstrated that low-hospital volume is associated with postoperative complications, extended length of stay, and inpatient mortality after surgery for primary brain tumors.<sup>12,14</sup> The high number of participating hospitals in the current study increases the generalizability of our findings; therefore, the results of this study may be more representative of typical management at all hospitals, including but not limited to tertiary care academic centers. Therefore, our results suggest a higher rate of postoperative major complications and mortality after craniotomy for a glioma than previous observational studies have suggested.<sup>15,30-33</sup>

To further investigate the generalizability of the current study, the demographics of our patient population were compared with those of the CBTRUS report on primary brain tumors (2009-2013) by Ostrom et al.<sup>1</sup> The age, sex, and race distributions in the current study paralleled the associated distributions stated in the CBTRUS report. These demographic similarities suggest that the study population of the current investigation is representative for primary malignant brain tumor patients across the US.<sup>1</sup>

Older age, comorbidity, and dependent functional status have previously been identified as predictors of postoperative complications in glioma patients,<sup>14,34,35</sup> and higher BMI as an additional predictor in brain tumor patients overall.<sup>7</sup> With regards to extended length of stay, older age and comorbidity have been identified as predictors in glioma patients,<sup>12,14</sup> and higher ASA-classification and dependent functional status as additional predictors in brain tumor patients in general.<sup>8</sup> For reoperations, higher BMI has been identified as a predictor in brain tumor patients.<sup>7</sup> For readmission, higher BMI, higher ASA-classification, dependent functional status, and steroid usage have been identified as predictors in brain tumor patients.<sup>4,7,29</sup> For mortality, older age and

comorbidity have been identified as predictors in glioma patients,<sup>11,12,14</sup> and higher ASA-classification, dependent functional status, steroid usage, emergency status, and longer operative time as additional predictors in brain tumor patients overall.<sup>7</sup>

**TABLE 5.** Multivariable logistic regression identifying predictors of death within 30 days after surgery.

<b>Predictor</b>	<b>Definition</b>	<b>Death</b>		
		<b>OR</b>	<b>95% CI</b>	<b>p</b>
Age	Per 10 years change	1.54 <sup>a</sup>	1.35-1.75	<.001
Gender	Female	Ref.	-	-
	Male	1.01	0.79-1.54	.56
BMI	Per 5 Kg/m <sup>2</sup> change	0.96 <sup>b</sup>	0.83-1.09	.52
ASA-classification	I-II	Ref.	-	-
	III	2.70	1.49-5.40	.002
	IV-V	5.95	3.11-12.37	<.001
Functional status	Independent	Ref.	-	-
	Dependent	2.46	1.52-3.83	<.001
WBC count	≤12,000/µL	Ref.	-	-
	>12,000/µL	1.58	1.12-2.21	.008
Model fit				
	AUC	0.76	0.72-0.80	<.001

ASA=American Society of Anesthesiologists, AUC: Area under the curve; BMI=Body Mass Index, CI=confidence interval; OR=odds ratio; p=p-value; WBC=white blood cell; For all tests, p<.0015 was considered significant.

<sup>a</sup>Inflated β-coefficients to odds ratio per 10 years increase.

<sup>b</sup>Inflated β-coefficients to odds ratio per 5 kg/m<sup>2</sup> increase.

To the best of our knowledge, this is the first study that uses the NSQIP registry to provide a comprehensive overview of the incidences and predictors of all clinically significant short-term outcomes after craniotomy focusing on the specific group primary malignant brain tumor patients. Additionally, it provides a time-to-event analysis for all complications, visualizes trends in complication rates over a seven-year time period, and investigates reasons for reoperation and readmission. Lastly, this study identifies novel predictors for short-term major complications (higher ASA-classification, higher BMI, elevated WBC, and longer operative time), increased length of stay (higher ASA-classification, dependent functional status, elevated WBC, low hematocrit, and ventilator dependence), reoperation (higher ASA-classification and elevated WBC), readmission (higher ASA-classification and dependent functional status) and mortality (higher ASA-classification and dependent functional status) after craniotomy for a primary malignant brain tumor.

## Implications

Possible reasons for poorer prognosis in elderly patients with primary malignant brain tumors likely include medical comorbidities and a lower overall fitness level.<sup>34,35</sup> Increased BMI also corresponds to a lower overall fitness level and has been demonstrated to be a risk factor of postoperative morbidity and mortality in other surgical specialties too.<sup>36</sup> ASA-classification is based on comorbidity of patients before surgery.<sup>37</sup> Although the relationship between comorbidity and postoperative complications is intuitive, the current study validates ASA-classification as a meaningful way to risk stratify glioma patients before surgery. Both ASA class III and ASA class IV-V were predictive for almost all outcome measures. An incremental association between ASA classification and postoperative unfavorable outcomes is suggested because the odds ratios were generally higher for ASA class IV-V compared to ASA class III. The strongest association was found between ASA class IV-V and death within 30 days (odds ratio 5.95). Functional dependence is often associated with underlying comorbidity or motor deficits, but it is also associated with poor rehabilitation after surgery; therefore, functional outcome impacts postoperative morbidity and even long-term survival.<sup>38</sup> After ASA-classification, functional dependence was the most frequent and strongest predictor of postoperative morbidity and mortality. For all outcome measures except reoperation, the odds were twice as high among functionally dependent patients. Increased WBC can indicate infection, inflammation, and malignancy.<sup>39,40</sup> In the current study, an association was found between elevated WBC and preoperative steroid usage ( $p<0.001$ ). Longer operative time results in a longer exposure to anesthesia and intraoperative risks; however, longer operative time also corresponds to surgical complexity, surgeon's experience, and other patient factors.<sup>41</sup> Although many of these factors have been described in previous literature as predictors of worse long-term outcomes in brain tumor or cancer patients in general, the current study also identifies them as predictors of short-term morbidity and mortality among patients operated on for a primary malignant brain tumor.

Most predictors are non-modifiable by surgeons; however, these results can help neurosurgeons and their multidisciplinary teams to identify high-risk patients for unfavorable outcomes after surgery. This may enable surgeons to tailor perioperative management to the risk profile of the individual patient. This is important because prophylactic treatment for one complication can increase the risk of other complications. For example, thromboprophylaxis can increase the risk of ICH. Optimizing the safety and efficacy of prophylactic strategies based on the risk profile of the individual patient can drastically reduce the rate of complications in the total population. Furthermore, targeting postoperative management can also reduce unnecessary healthcare costs.

## **Reoperation**

Reoperation also qualifies as a major complication and was the most common major complication in this study.<sup>27</sup> Reoperation is an important indicator of worse clinical outcome recorded in NSQIP and inherently involves increased costs and risks to patients. In this study, ICH was found to be the primary reason for reoperation (18.5% of all patients reoperated within 30 days). Postoperative hemorrhage is one of the most serious complications of any operation on the brain, and is associated with significant morbidity and mortality in addition to that from the original operation and the primary disease.<sup>42,43</sup> ICH is difficult to define and may include bleeding following craniotomy at the operative site or remotely, though this is rare.<sup>44-46</sup> The rates of postoperative ICH following intracranial operations vary greatly throughout the existing literature (0.8-50%).<sup>42,47,48</sup> Hypertension and decreased factor XIII have been identified as factors associated with ICH after brain tumor surgery.<sup>49,50</sup> Interestingly, resection of residual tumor tissue was identified as the third most common cause of reoperation. Improving the extent of resection has gained more attention in recent years, and many complex modalities have been developed and applied intraoperatively to guide and monitor surgical resection, such as stereotactic navigation, intraoperative MRI, ultrasound, functional mapping, and fluorescence guided surgery.<sup>51</sup> The use of these modalities is highly dependent on their availability and the surgeon's preference; however, they can potentially reduce the rate of short-term reoperations as they find their way to standard clinical practice.

## **Readmission**

Readmission is a major driver of cost and re-exposes patients to associated risks of long hospital stays.<sup>52,53</sup> The most common causes for readmission following craniotomy for glioma resection were found to be wound related, occurring in 11.9% of readmitted patients. Issues with wound healing, including infection, are known and commonly reported complications of brain tumor resections.<sup>11,15,54,55</sup> Risk factors for wound-related complications include having previously undergone additional craniotomies, additional radiosurgery, or having been treated with the anti-angiogenic factor bevacizumab.<sup>54,56,57</sup>

## **Limitations**

Limitations of this study are primarily a result of variables not included in the NSQIP dataset, potentially causing underestimation of the total complication and mortality rates for craniotomies. Tumor and surgery specific information, such as histology, grading, size, and location of the tumor as well as the extent of resection have an enormous impact on both short and long-term outcomes; however, these are not

available in the NSQIP registry. This can cause uncontrollable confounding, as is also demonstrated in a previous study.<sup>58</sup> Underestimation in this case might also be caused by underreporting and selection bias as participating hospitals are not obliged to contribute all consecutive cases; however, data is collected on randomly assigned patients by trained surgical reviewers, and inter-rater reliability audits are performed to ensure data reliability. Complications after the 30-day limit used by the database are also unaccounted for in this study. The breadth of multicenter data from NSQIP used in the current study is more representative than most single-center reports; however, the effects of surgeon experience or center volume on postoperative outcomes cannot be accounted for in this database. All surgical studies are limited by variability in surgeon experience,<sup>59</sup> as well as geographical location,<sup>1</sup> both of which have been demonstrated to independently affect complication rates in neurosurgery.

Despite these limitations, this study provides useful insight into the rates, reasons, timing, and predictors of major complications, extended length of stay, reoperation, readmission, and mortality after craniotomy for primary malignant brain tumors. Future studies should focus on building advanced prediction models for short-term outcomes after craniotomy, enabling physicians to tailor postoperative management to the risk profile of the individual patient. A national neurosurgical quality improvement registry including tumor specific and neurosurgical variables can be essential for achieving this goal.

## **Conclusion**

Among patients undergoing craniotomy for primary malignant brain tumors, 12.9% experienced a major complication within 30 days after surgery, most of which occurred during the initial hospital stay. Intracranial hemorrhage and wound-related complications were the major causes of reoperation and readmission, respectively. ASA-classification and dependent functional status are primarily predictive for morbidity and mortality within 30 days after craniotomy for a primary malignant brain tumor. Future inclusion of tumor and neurosurgical specific variables could allow for a more granular risk assessment of short-term outcomes after craniotomy, but the lack of these variables currently limits the implications of this study.

## References

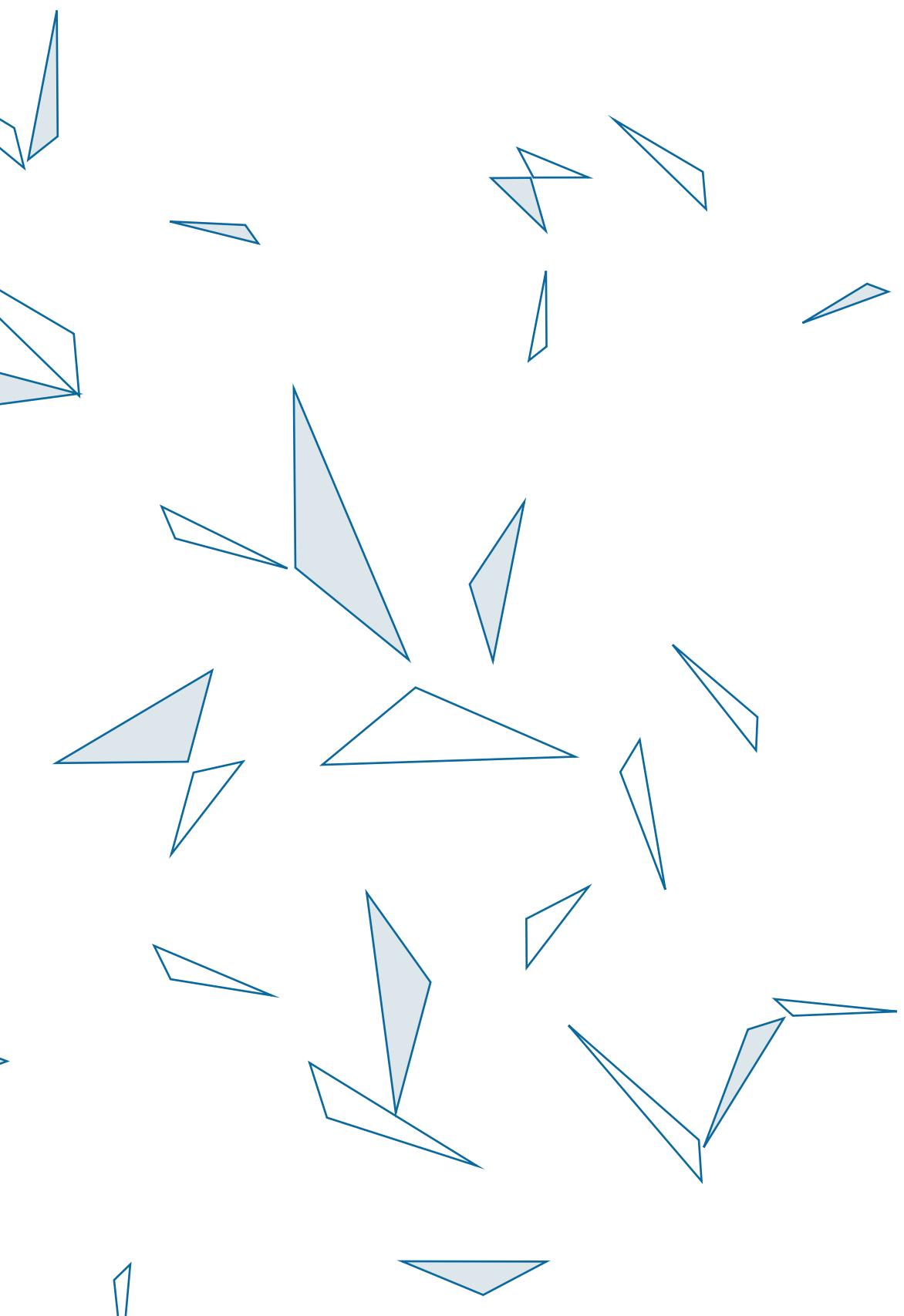
1. Ostrom QT GH, Xu J, Kromer C, Wolinsky Y, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2009-2013. *Neuro-oncology*. 2016;18:v1-v75.
2. Stupp R HM, Gilbert MR, Chakravarti A. Chemoradiotherapy in malignant glioma: standard of care and future directions. *Journal of Clinical Oncology*. 2007;25(26):4127-4136.
3. Krivosheya D PS, Weinberg JS, Sawaya R. Technical principles in glioma surgery and preoperative considerations. *Journal of Neuro-oncology*. 2016;130(2):243-252.
4. Alan N, Seicean A, Seicean S, Neuhauser D, Benzel EC, Weil RJ. Preoperative steroid use and the incidence of perioperative complications in patients undergoing craniotomy for definitive resection of a malignant brain tumor. *J Clin Neurosci*. 2015;22(9):1413-1419.
5. Chang SM, Parney IF, McDermott M, et al. Perioperative complications and neurological outcomes of first and second craniotomies among patients enrolled in the Glioma Outcome Project. *Journal of neurosurgery*. 2003;98(6):1175-1181.
6. Dasenbrock HH, Devine CA, Liu KX, et al. Thrombocytopenia and craniotomy for tumor: A National Surgical Quality Improvement Program analysis. *Cancer*. 2016.
7. Dasenbrock HH, Liu KX, Chavakula V, et al. Body habitus, serum albumin, and the outcomes after craniotomy for tumor: a National Surgical Quality Improvement Program analysis. *Journal of neurosurgery*. 2017;126(3):677-689.
8. Dasenbrock HH, Liu KX, Devine CA, et al. Length of hospital stay after craniotomy for tumor: a National Surgical Quality Improvement Program analysis. *Neurosurgical focus*. 2015;39(6):E12.
9. Lassen B, Helseth E, Ronning P, et al. Surgical mortality at 30 days and complications leading to reoperation in 2630 consecutive craniotomies for intracranial tumors. *Neurosurgery*. 2011;68(5):1259-1268; discussion 1268-1259.
10. Marcus LP, McCutcheon BA, Noorbakhsh A, et al. Incidence and predictors of 30-day readmission for patients discharged home after craniotomy for malignant supratentorial tumors in California (1995-2010). *Journal of neurosurgery*. 2014;120(5):1201-1211.
11. Missios S KP, Nanda A, Bekelis K. Craniotomy for glioma resection: a predictive model. *World neurosurgery*. 2015;83(6):957-964.
12. Nuno M, Mukherjee D, Carico C, et al. The effect of centralization of caseload for primary brain tumor surgeries: trends from 2001-2007. *Acta Neurochir (Wien)*. 2012;154(8):1343-1350.
13. Seicean A, Seicean S, Schiltz NK, et al. Short-term outcomes of craniotomy for malignant brain tumors in the elderly. *Cancer*. 2013;119(5):1058-1064.
14. Trinh VT, Davies JM, Berger MS. Surgery for primary supratentorial brain tumors in the United States, 2000-2009: effect of provider and hospital caseload on complication rates. *Journal of neurosurgery*. 2015;122(2):280-296.
15. Chang SM PI, McDermott M, Barker FG, Schmidt MH, Huang W, Laws Jr ER, Lillehei KO, Bernstein M, Brem H, Sloan AE, Berger M, Glioma Outcomes Investigators. Perioperative complications and neurological outcomes of first and second craniotomies among patients enrolled in the Glioma Outcome Project. *Journal of neurosurgery*. 2003;98(6):1175-1181.
16. ACS NSQIP Hospitals. 2017; <https://www.facs.org/search/nsqip-participants?allresults=>. Accessed October 9, 2017.

17. Sellers MM, Merkow RP, Halverson A, et al. Validation of new readmission data in the American College of Surgeons National Surgical Quality Improvement Program. *Journal of the American College of Surgeons.* 2013;216(3):420-427.
18. Lieber BA, Appelboom G, Taylor BE, Malone H, Agarwal N, Connolly ES, Jr. Assessment of the "July Effect": outcomes after early resident transition in adult neurosurgery. *Journal of neurosurgery.* 2015;1-9.
19. McGirt MJ, Godil SS, Asher AL, Parker SL, Devin CJ. Quality analysis of anterior cervical discectomy and fusion in the outpatient versus inpatient setting: analysis of 7288 patients from the NSQIP database. *Neurosurgical focus.* 2015;39(6):E9.
20. Lieber BA, Appelboom G, Taylor BE, et al. Preoperative chemotherapy and corticosteroids: independent predictors of cranial surgical-site infections. *Journal of neurosurgery.* 2015;1-9.
21. Lim S, Parsa AT, Kim BD, Rosenow JM, Kim JY. Impact of resident involvement in neurosurgery: an analysis of 8748 patients from the 2011 American College of Surgeons National Surgical Quality Improvement Program database. *Journal of neurosurgery.* 2015;122(4):962-970.
22. McCutcheon BA, Ciacci JD, Marcus LP, et al. Thirty-Day Perioperative Outcomes in Spinal Fusion by Specialty Within the NSQIP Database. *Spine.* 2015;40(14):1122-1131.
23. Lieber BA, Han J, Appelboom G, et al. Association of Steroid Use with Deep Venous Thrombosis and Pulmonary Embolism in Neurosurgical Patients: A National Database Analysis. *World neurosurgery.* 2016.
24. Kim BD, Smith TR, Lim S, Cybulski GR, Kim JY. Predictors of unplanned readmission in patients undergoing lumbar decompression: multi-institutional analysis of 7016 patients. *Journal of neurosurgery Spine.* 2014;20(6):606-616.
25. Hackett NJ, De Oliveira GS, Jain UK, Kim JY. ASA class is a reliable independent predictor of medical complications and mortality following surgery. *International journal of surgery (London, England).* 2015;18:184-190.
26. Abt NB, De la Garza-Ramos R, Olorundare IO, et al. Thirty day postoperative outcomes following anterior lumbar interbody fusion using the national surgical quality improvement program database. *Clinical neurology and neurosurgery.* 2016;143:126-131.
27. Lukasiewicz AM, Grant RA, Basques BA, Webb ML, Samuel AM, Grauer JN. Patient factors associated with 30-day morbidity, mortality, and length of stay after surgery for subdural hematoma: a study of the American College of Surgeons National Surgical Quality Improvement Program. *Journal of neurosurgery.* 2016;124(3):760-766.
28. Dasenbrock HH, Devine CA, Liu KX, et al. Thrombocytopenia and craniotomy for tumor: A National Surgical Quality Improvement Program analysis. *Cancer.* 2016;122(11):1708-1717.
29. Dasenbrock HH, Yan SC, Smith TR, et al. Readmission After Craniotomy for Tumor: A National Surgical Quality Improvement Program Analysis. *Neurosurgery.* 2017;80(4):551-562.
30. Chaichana KL MM, Laterra J, Olivi A, Quinones-Hinojosa A. Recurrence and malignant degeneration after resection of adult hemispheric low-grade gliomas. *Journal of neurosurgery.* 2010;112(1):10-17.
31. Stark AM HJ, Held-Feindt J, Mehdorn HM. Glioblastoma-the consequences of advanced patient age on treatment and survival. *Neurosurgical Review.* 2007;30(1):56-61.
32. Cric I AM, Vick N, Mikhael M. Supratentorial gliomas: surgical considerations and immediate postoperative results. *Neurosurgery.* 1987;21(1):21-26.
33. Devaux BC OFJ, Kelly PJ. Resection, biopsy, and survival in malignant glial neoplasms. A retrospective study of clinical parameters, therapy, and outcome. *Journal of neurosurgery.* 1993;78(5):767-775.
34. Cric I, Ammirati M, Vick N, Mikhael M. Supratentorial gliomas: surgical considerations and immediate postoperative results. Gross total resection versus partial resection. *Neurosurgery.* 1987;21(1):21-26.

35. D'Amico RS, Cloney MB, Sonabend AM, et al. The Safety of Surgery in Elderly Patients with Primary and Recurrent Glioblastoma. *World neurosurgery*. 2015;84(4):913-919.
36. Turrentine FE, Hanks JB, Schirmer BD, Stukenborg GJ. The relationship between body mass index and 30-day mortality risk, by principal surgical procedure. *Arch Surg*. 2012;147(3):236-242.
37. Wolters U, Wolf T, Stutzer H, Schroder T. ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth*. 1996;77(2):217-222.
38. Walid MS. Prognostic factors for long-term survival after glioblastoma. *Perm J*. 2008;12(4):45-48.
39. Haim M, Boyko V, Goldbourt U, Battler A, Behar S. Predictive value of elevated white blood cell count in patients with preexisting coronary heart disease: the Bezafibrate Infarction Prevention Study. *Arch Intern Med*. 2004;164(4):433-439.
40. Zadora P, Dabrowski W, Czarko K, et al. Preoperative neutrophil-lymphocyte count ratio helps predict the grade of glial tumor - a pilot study. *Neurol Neurochir Pol*. 2015;49(1):41-44.
41. Jackson TD, Wannares JJ, Lancaster RT, Rattner DW, Hutter MM. Does speed matter? The impact of operative time on outcome in laparoscopic surgery. *Surg Endosc*. 2011;25(7):2288-2295.
42. Seifman MA LP, Rosenfeld JV, Hwang PYK. Postoperative intracranial haemorrhage: a review. *Neurosurgical Review*. 2011;34(4):393-407.
43. Fadul C WJ, Thaler H, Galichich J, Patterson Jr RH, Posner JB. Morbidity and mortality of craniotomy for excision of supratentorial gliomas. *Neurology*. 1988;38:1374-1379.
44. Borkar SA LG, Sharma BS, Mahapatra AK. Remote site intracranial hemorrhage: a clinical series of five patients with review of literature. *British Journal of Neurosurgery*. 2013;27(6):735-738.
45. Garg KTV, Sinha S, Suri A, Mahapatra AK, Sharma BS. Remote site intracranial hemorrhage: our experience and review of the literature. *Neurology India*. 2014;62(3):296-302.
46. Brisman MH BJ, Sen CN, Germano IM, Moor F, Post KD. Intracerebral hemorrhage occurring remote from the craniotomy site. *Neurosurgery*. 1996;39(6):1114-1121.
47. Fukamachi A KH, Nukui H. Postoperative intracerebral hemorrhages: a survey of computed tomographic findings after 1074 intracranial operations. *Surgical Neurology*. 1985;23(6):575-580.
48. Kalfas IH LJ. Postoperative hemorrhage: a survey of 4992 intracranial procedures. *Neurosurgery*. 1988;23(3):343-347.
49. Gerlach R, Raabe A, Zimmermann M, Siegemund A, Seifert V. Factor XIII deficiency and postoperative hemorrhage after neurosurgical procedures. *Surg Neurol*. 2000;54(3):260-264; discussion 264-265.
50. Kalfas IH, Little JR. Postoperative hemorrhage: a survey of 4992 intracranial procedures. *Neurosurgery*. 1988;23(3):343-347.
51. Roder C, Bisdas S, Ebner FH, et al. Maximizing the extent of resection and survival benefit of patients in glioblastoma surgery: high-field iMRI versus conventional and 5-ALA-assisted surgery. *Eur J Surg Oncol*. 2014;40(3):297-304.
52. Friedman B BJ. The rate and cost of hospital readmissions for preventable conditions. *Medical Care Research and Review*. 2004;61(2):225-240.
53. Chetty VK CL, Phillips RL Jr, Rankin J, Xierali I, Finnegan S, Jack B. FPs lower hospital readmission rates and costs. *American Family Physician*. 2011;83(9):1054.
54. Barami K FR. Incidence, risk factors and management of delayed wound dehiscence after craniotomy for tumor resection. *Journal of Clinical Neuroscience*. 2012;19(6):854-857.
55. Cabantog AM BM. Complications of first craniotomy for intra-axial brain tumour. *The Canadian journal of neurological sciences*. 1994;21(3):213-218.

56. Ladha H PT, Gilbert MR, Mandel J, O'Brien B, Conrad C, Fields M, Hanna T, Loch C, Armstrong TS. Wound healing complications in brain tumor patients on Bevacizumab. *Journal of Neuro-oncology*. 2015;124(3):501-506.
57. Clark AJ BN, Chang SM, Prados MD, Clarke J, Polley MY, Sughrue ME, McDermott MW, Parsa AT, Berger MS, Aghi MK. Impact of bevacizumab chemotherapy on craniotomy wound healing. *Journal of neurosurgery*. 2011;114(6):1609-1616.
58. Grabowski MM RP, Nowacki AS< Schroeder JL, Angelov L, Barnett GH, Vogelbaum MA. Residual tumor volume versus extent of resection: predictors of survival after surgery for glioblastoma. *Journal of neurosurgery*. 2014;121(5):1115-1123.
59. Cricic I RA, Baumgartner C, Pierce D. Complications of transsphenoidal surgery: results of a national survey, review of the literature, and personal experience. *Neurosurgery*. 1997;40(2):225-236.





# 3

## Venous thromboembolism and intracranial hemorrhage after craniotomy for primary malignant brain tumors: A National Surgical Quality Improvement Program analysis

---

Joeky T. Senders, Nicole H. Goldhaber, David J. Cote, Ivo S. Muskens, Hassan Y. Dawood, Filip Y.F.L. De Vos, William B. Gormley, Timothy R. Smith\*, Marike L.D. Broekman\*

\*Shared last authorship

J NEUROONCOL. 2018 JAN;136(1):135-145

## **Abstract**

Venous thromboembolism (VTE), including deep venous thrombosis (DVT) and pulmonary embolism (PE), frequently complicates the postoperative course of primary malignant brain tumor patients. Thromboprophylactic anticoagulation is commonly used to prevent VTE at the risk of intracranial hemorrhage (ICH). We extracted all patients who underwent craniotomy for a primary malignant brain tumor from the National Surgical Quality Improvement Program (NSQIP) registry (2005-2015) to perform a descriptive time-to-event analysis and identify relevant predictors of DVT, PE, and ICH within 30 days after surgery. Among the 7376 identified patients, the complication rates were 2.6%, 1.5%, and 1.3% for DVT, PE, and ICH, respectively. VTE was the second-most common major complication and third-most common reason for readmission. ICH was the most common reason for reoperation. The increased risk of VTE extends beyond the period of hospitalization, especially for PE, whereas ICH occurred predominantly within the first days after surgery. Older age and higher BMI were confirmed as overall predictors of VTE. Dependent functional status and longer operative times were predictive for VTE during hospitalization, but not for post-discharge events. Admission two or more days before surgery was predictive for DVT, but not for PE. Preoperative steroid usage and male gender were predictive for post-discharge DVT and PE, respectively. ICH was associated with various comorbidities and longer operative times. This multicenter study demonstrated distinct critical time periods for the development of thrombotic and hemorrhagic events after craniotomy. Furthermore, the VTE risk profile depends on the type of VTE (DVT versus PE) and clinical setting (during hospitalization versus after discharge).

## Introduction

Venous thromboembolism (VTE), including deep venous thrombosis (DVT) and pulmonary embolism (PE), constitutes a major cause of morbidity and mortality in patients undergoing craniotomy for a primary malignant brain tumor.<sup>1-6</sup> Cancer is a recognized risk factor for VTE development in addition to known surgical risk factors, such as venous stasis from perioperative immobility, endothelial injury, and inflammation from the operation itself.<sup>7</sup> Among all cancer types, high-grade gliomas have been shown to result in the second highest lifetime risk for cancer-related VTE and one of the highest risks of perioperative VTE. Rates of postoperative VTE have been reported to be twice as high when comparing craniotomy for any brain tumor versus non-neoplastic diseases.<sup>8-11</sup> VTE has been reported as one of the most frequent major complications after craniotomy for brain tumors with incidences up to 21% in the first three months after surgery.<sup>1-6</sup>

Previous studies have identified older age, male sex, Hispanic ethnicity, history of craniotomy, history of VTE, congestive heart failure, coagulopathy, seizures, increased stay on the intensive care unit, prolonged hospital stay, residual tumor tissue, and absence of thromboprophylactic therapy as predictors of VTE after craniotomy for a primary malignant brain tumor.<sup>3,4,6,12-15</sup> Most of these studies have been small, single-center studies, and none of these studies have identified predictors or performed time-to-event analyses stratified for VTE type (DVT versus PE) or clinical setting (during hospitalization versus after discharge).

Most patients undergoing surgery for a brain tumor receive pharmaceutic prophylaxis in combination with mechanical prophylaxis in the perioperative setting.<sup>16-19</sup> However, anticoagulation increases the risk of intracranial hemorrhage (ICH), which is one of the most frequent and feared complications in patients undergoing brain tumor surgery.<sup>20</sup> The increased risk of ICH makes the use of prophylactic anticoagulation an issue of great debate and careful balance in this patient population. Although the incidence of ICH is lower compared to VTE events, their outcomes can be at least as detrimental. Only few predictors associated with ICH have been identified including history of craniotomy, use of bevacizumab, and therapeutic anticoagulation for a VTE.<sup>13,21-25</sup> Adequate assessment of the perioperative risk of both VTE and ICH within this patient population, as well as accurately characterizing the timing of thrombotic and hemorrhagic events, is meaningful for tailoring postoperative management to the risk profile of the individual patient.

The American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) database tracks outcomes of neurosurgical patients for 30 days postoperatively, including the occurrence of DVT, PE, and ICH. Because the risk of ICH is intimately tied to the thromboprophylactic strategy used in patients with a brain tumor, this study addresses thrombotic as well as hemorrhagic complications. In this study, NSQIP was used to identify predictors and perform descriptive time-to-event analyses for VTE and ICH. Assessment of VTE was stratified for VTE type (DVT versus PE) and clinical setting (during hospitalization versus after discharge).

## Methods

### Data source

The NSQIP includes surgical patients from over 600 participating hospitals in the U.S. operated on from 2005 to 2015. This validated dataset is collected by trained surgical reviewers using a standardized protocol, and includes common postoperative complications, occurrence of reoperations and readmissions, together with associated reasons and time-to-event in days. The NSQIP registry has previously been used to study outcomes after neurosurgical procedures.<sup>26-37</sup> Our institutional review board has exempted the NSQIP database from review.

### Inclusion criteria

Patients were included who met the following criteria: 1) age 18 years or older; 2) a Current Procedural Terminology (CPT) code indicating craniotomy for surgical resection of a brain tumor (CPT: 61500, 61510, 61512, 61518, 61519, 61520, 61521, 61526, and 61530); 3) a postoperative diagnosis indicative of a primary malignant brain tumor (International Classification of Diseases, Ninth Revision [ICD-9]: 191.x).

### Covariates

Age, body mass index (BMI), and operative time were assessed continuously in years, kg/m<sup>2</sup>, and minutes, respectively. Other categorized pre- and perioperative covariates included sex, race, American Society of Anesthesiologists (ASA)-classification (I-II, III or IV-V), functional status (dependent or independent), smoking within one year prior to surgery, history of hypertension requiring medication, chronic heart failure, COPD, renal failure, dialysis, insulin dependent diabetes, bleeding disorder, weight loss (>10% loss of body weight in the six months prior to surgery), dyspnea, ventilator dependence, steroid usage, emergency classification, transfer status (admitted from home versus elsewhere), creatinine (<1.4 mg/dL vs. ≥1.4 mg/dL), hematocrit (<36% vs.

≥36%), platelet count (100-450/ $\mu\text{m}^3$ , <100/ $\mu\text{m}^3$ , vs. >450/ $\mu\text{m}^3$ ), sodium (135-145 mEq/L, <135 mEq/L, vs. >145 mEq/L), white blood cell (WBC) count ( $\leq$ 12,000/ $\mu\text{L}$  vs. >12,000/ $\mu\text{L}$ ), preoperative transfusion, preoperative systemic inflammatory response syndrome, admission two or more days before surgery, and anesthesia type (general versus no general anesthesia).

## **Missing data**

Covariates with more than ten percent missing data or occurring in less than one percent of cases were excluded from multivariable analysis. Cases with missing data in one of the variables of the multivariable analysis were excluded from the analysis. A confirmatory analysis was performed for every multivariable model, in which missing data was coded as an additional group to verify if missing data affected the results.

## **Outcomes**

VTE was defined as the occurrence of DVT or PE within 30 days after surgery. The occurrence of other major complications and reasons for readmission and reoperation were extracted by means of ICD-9 and CPT codes, to assess the relative contribution of VTE and ICH to morbidity, readmission, and reoperation. Based on a previously published definition, major complications were defined as either acute renal failure, cardiac arrest, death, failure to wean from ventilator, myocardial infarction, reintubation, reoperation, stroke, VTE, sepsis, and surgical site infection.<sup>38</sup> ICH was defined as the occurrence of an ICH requiring surgical evacuation and extracted by means of CPT and ICD-9 codes. Reasons for reoperation including ICH were collected since 2012. Each primary thrombotic outcome (VTE, DVT, PE) was assessed for its occurrence during the initial hospital stay, after discharge, and within 30 days overall. ICH was assessed within 30 days overall.

## **Statistical analysis**

Statistical analyses were conducted using R 3.3.3 (R Core Team, Auckland, New Zealand). Univariable analysis was performed using logistic regression. Potential predictors were selected for inclusion in the multivariable logistic regression analysis based on univariable analysis for each outcome. Only pre- and intraoperative factors were included in the multivariable analysis because inclusion of postoperative complications other than VTE or ICH would reduce the timeframe in which complications can be detected due to the limited 30-day collection period of NSQIP, thereby biasing the results of the model. Potential predictors were excluded from the final model if they demonstrated multicollinearity or had a relative low contribution to model fit.

A p-value below 0.05 was considered statistically significant. Bonferroni correction for multiple testing was deemed to be too rigorous due to the low number of events. The  $\beta$ -coefficients of the continuous variables in the final model were multiplied to represent the odds ratios and confidence intervals of meaningful and interpretable units for age (per ten years increase), BMI (per five kg/m<sup>2</sup> increase), and operative time (per 60 minutes increase).

## Results

### Demographics of study population

The NSQIP registry provided 7376 patients who underwent craniotomy for resection of a primary malignant brain tumor during the study period. Comorbidities, demographics, and preoperative laboratory values are shown separated by the occurrence of VTE (Table 1).

### Outcomes

Of the 7376 patients that were identified, 257 (3.5%) developed a VTE within 30 days after surgery, of which 91 (36%) occurred within the initial hospital stay. VTE was the second-most common major complication and included 192 DVTs (2.6%) and 107 PEs (1.5%). Forty-two patients developed both DVT and PE (0.6%). The rate of DVT was highest within the first two weeks after surgery, whereas the rate of PE was fairly consistent throughout the first month, occurring predominantly post-discharge (Fig. 1a, Fig. 1b). The rate of VTE was more than twice as high (7.0% versus 3.2%,  $p < 0.001$ ) among patients with a preoperative dependent functional status compared to patients with an independent functional status (Fig. 1c).

The median length of stay among VTE patients was eight days (inter-quartile range [IQR] 5-16 days) compared to four days (IQR 3-8 days) in non-VTE patients (Fig. 1d). In-hospital VTE occurred at a median of six days (IQR 3-8 days) after surgery, and patients were discharged at a median of eight days after the occurrence of VTE (IQR 3-16 days). Post-discharge VTE occurred at a median of 13 days after discharge (IQR 6-19 days) and resulted in 25% of cases in readmission, making VTE the third-most common reason for readmission (7.4% of total readmissions). Of the patients that developed a PE, 35 (32.7%) were preceded by a DVT. Post-discharge PEs were less frequently preceded by a DVT than hospital acquired PEs (26% vs. 48%,  $p=0.048$ ).

**TABLE 1.** Demographics and preoperative comorbidities of NSQIP patients undergoing craniotomy for glioma, compared by VTE occurrence.

Characteristic (%)	Definition	Total (n = 7376)	No VTE (n = 7119)	VTE (n = 257)	OR	95% CI	p
Age	Years±SD	54.5±15.6	54.4±15.6	59.4±15.7	1.25 <sup>a</sup>	1.15-1.36	<0.001
Gender							
Female	42.3	42.3	41.6	Ref.	-	-	-
Male	57.7	57.7	58.4	1.03	0.91-1.19	0.82	
Race							
White	91.7	91.7	94.8	Ref.	-	-	-
Black	4.8	4.8	4.2	0.85	0.40-1.60	0.66	
Asian	3.0	3.0	0.5	0.15	0.01-0.67	0.04	
Other	0.5	0.5	0.5	0.95	0.05-4.51	0.96	
ASA-classification							
I-II	27.7	28.1	18.3	Ref.	-	-	-
III	53.2	59.0	65.5	1.71	1.23-2.40	0.002	
IV-V	13.0	12.9	16.3	1.94	1.26-2.97	0.002	
BMI	kg/m <sup>2</sup> ±SD	28.4±6.2	28.4±6.2	29.8±6.3	1.19 <sup>b</sup>	1.08-1.30	<0.001
Smoking							
Never	17.1	17.4	10.1	0.53	0.35-0.78	0.003	-
Ever	6.5	6.5	8.9	1.42	0.89-2.15	0.12	
Emergency case							
Admitted not from home							
Hypertension							
Admitted not from home							
History of COPD							
History of CHF							
Renal failure							
Dialysis							
Ventilator dependence							
Weight loss							
Bleeding disorder							
Dyspnea							

TABLE 1. Continued

Characteristic (%)	Definition	Total (n = 7376)	No VTE (n = 7119)	VTE (n = 257)	OR	95% CI	p
Insulin-dependent DM	3.9	3.8	4.7	1.23	0.65-2.13	0.49	
Pre-op steroid usage	16.6	16.4	23.3	1.55	1.15-2.07	0.004	
Functional dependence	5.1	5.0	10.4	2.23	1.43-3.32	<0.001	
Pre-op SIRS	3.6	3.4	5.1	1.51	0.81-2.56	0.16	
Pre-op transfusion	0.2	0.2	0.4	1.85	0.10-9.18	0.55	
Pre-op Sodium	135-145	89.9	87.2	Ref.	-	-	
	<135	9.1	9.0	10.8	1.24	0.80-1.83	0.31
	>145	1.0	1.0	2.0	2.07	0.72-4.69	0.12
Pre-op creatinine	≥ 1.4 mg/dL	4.4	4.6	4.4	0.96	0.49-1.69	0.90
Pre-op WBC	>12000/ $\mu$ L	24.8	24.4	33.1	1.53	1.16-1.99	0.002
Pre-op hematocrit	<36%	11.9	11.8	12.6	1.07	0.72-1.54	0.71
Platelets	100-450	97.5	97.6	95.7	Ref.	-	-
	<100	1.3	1.0	2.8	2.29	0.95-4.65	0.05
	>450	1.2	1.2	1.6	1.37	0.41-3.33	0.54
Operative time	min [IQR]	179[123-250]	179[123-250]	191[134-252]	1.33 <sup>c</sup>	1.02-1.74	0.04
No general anesthesia		5.9	5.9	5.4	0.91	0.50-1.52	0.74
Admission to operation	≥ 2 days	32.8	32.4	46.3	1.80	1.40-2.31	<0.001

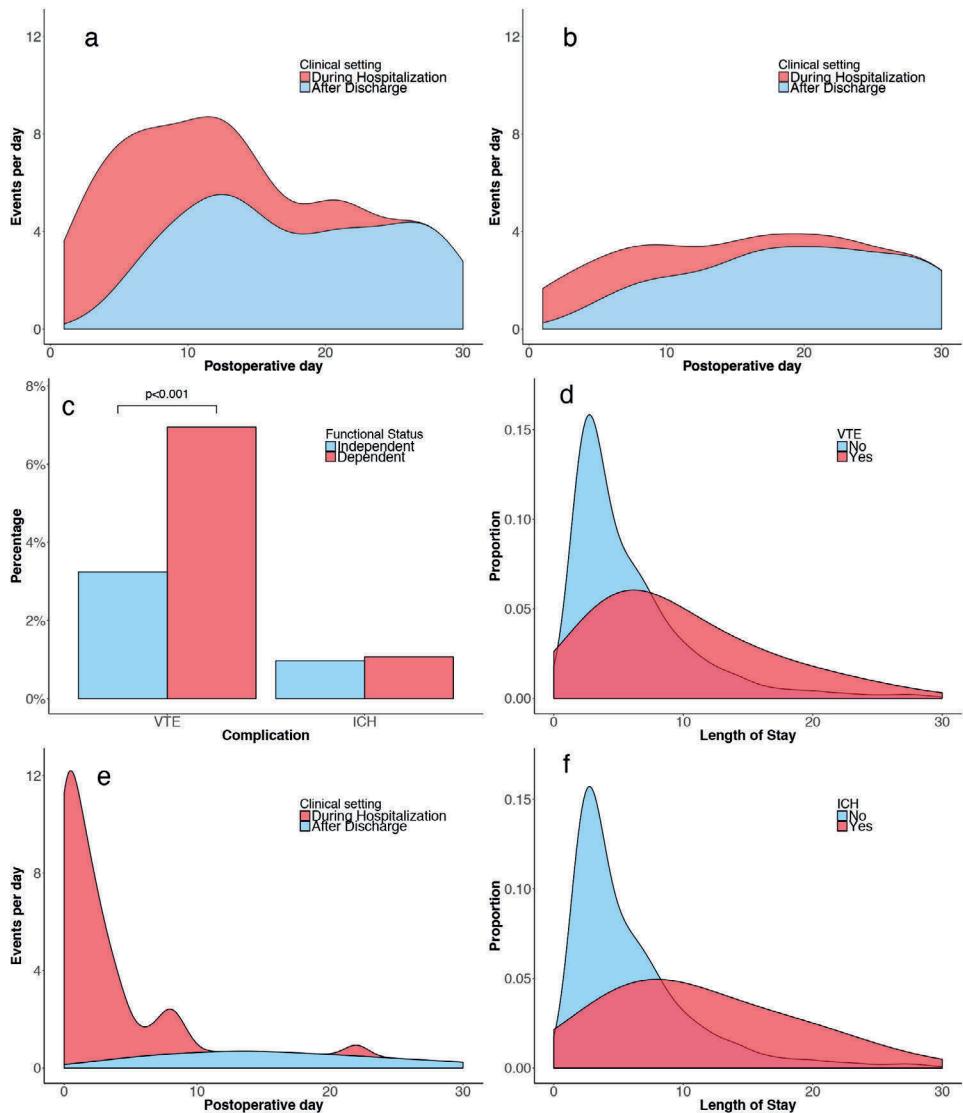
Abbreviations: ASA=American Society of Anesthesiologists, CI=confidence interval, CHF=congestive heart failure, p=p-value; pre-op=preoperative; SIRS=systematic inflammatory response syndrome, WBC=white blood cell count.

<sup>a</sup>Inflated β-coefficients to odds ratio per 10 years increase

<sup>b</sup>Inflated β-coefficients to odds ratio per 5 kg/m<sup>2</sup> increase

<sup>c</sup>Inflated β-coefficients to odds ratio per 60 minutes increase

<sup>d</sup>Infinity due to 0 count in one of the cells



**FIGURE 1.** Number of patients per day in the total population developing a deep venous thrombosis (a), pulmonary embolism (b) or intracranial hemorrhage (e) after craniotomy stratified for clinical setting. Distribution of length of postoperative stay compared by the occurrence of VTE (d) and ICH (f). Frequency of VTE and ICH compared by functional status (c).

Of the 5699 patients that were identified in the NSQIP registry 2012-2015 with data on reoperation and associated reasons, 72 (1.3%) developed an ICH requiring surgical evacuation at a median of two days after surgery (IQR 0-8 days) (Fig 1e). ICH was the most common reason for reoperation (18.5% of the total number of reoperations). The median length of stay among ICH patients was 12 days (IQR 6-21 days) compared to four days (IQR 3-8 days) in non-ICH patients (Fig. 1f). The 55 patients (77.5%) that developed an ICH during the initial hospital stay, were discharged at a median of ten days (IQR 6-19) after the occurrence of ICH.

**TABLE 2.** Multivariable analysis comparing risk profiles for deep venous thrombosis and pulmonary embolism occurring in-hospital versus post-discharge.

<b>Predictors</b>	<b>In-hospital DVT</b>			<b>Post-discharge DVT</b>		
	<b>OR</b>	<b>95% CI</b>	<b>p</b>	<b>OR</b>	<b>95% CI</b>	<b>p</b>
Age per 10 years increase	1.28	1.06-1.55	0.01	1.28	1.12-1.47	<0.001
BMI per 5 kg/m <sup>2</sup> increase	1.23	1.01-1.46	0.03	1.25	1.09-1.41	<0.001
Dependent functional status	2.63	1.18-5.27	0.01	0.82	0.34-1.68	0.62
Preoperative WBC >12000/ $\mu$ L	1.76	1.00-2.97	0.05	1.05	0.69-1.57	0.80
Steroid usage	1.38	0.71-2.52	0.32	2.17	1.39-3.30	<0.001
Admission $\geq$ 2 days before OR	1.85	1.08-3.19	0.02	2.02	1.37-2.98	<0.001
OR time per 60 min increase	1.26	1.13-1.39	<0.001	1.04	0.94-1.15	0.40
<hr/>						
<b>Predictors</b>	<b>In-hospital PE</b>			<b>Post-discharge PE</b>		
	<b>OR</b>	<b>95% CI</b>	<b>p</b>	<b>OR</b>	<b>95% CI</b>	<b>p</b>
Age per 10 years increase	1.08	0.84-1.41	0.56	1.33	1.13-1.58	<0.001
Male gender	0.93	0.43-2.04	0.85	2.32	1.38-4.07	0.002
BMI per 5 kg/m <sup>2</sup> increase	1.27	0.97-1.62	0.07	1.24	1.04-1.46	0.01
Dependent functional status	5.46	1.96-13.12	<0.001	2.15	0.94-4.31	0.05
OR time per 60 min increase	1.24	1.06-1.43	0.004	1.10	0.98-1.23	0.10

Abbreviations: BMI=body mass index; DVT=deep venous thrombosis; min=minutes; OR=operation room; PE=pulmonary embolism; WBC=white blood cell count.

## Multivariable analysis

Older age and higher BMI were found to be risk factors of VTE overall (Tables 2-3). Dependent functional status and longer operative times were predictive for hospital VTE, but not for post-discharge events. Admission two or more days before surgery was a predictor of DVT, but not for PE. Steroid usage was predictive for post-discharge DVT, and male gender was predictive for post-discharge PE. Higher ASA-classification,

hypertension, weight loss, bleeding disorders, preoperative sodium < 135 mEq/L, and longer operative times were found to be predictors of postoperative ICH requiring surgical evacuation (Table 4).

**TABLE 3.** Overview overall and specific risk factors for VTE.

Overall risk factors		
	In-Hospital	Post-Discharge
DVT	- Admission ≥2 days pre-op - Longer operative times - Dependent functional status	- Admission ≥2 days pre-op - Steroid usage
PE <sup>a</sup>	- Longer operative times - Dependent functional status	- Male gender

Abbreviations: BMI=body mass index; DVT=deep venous thrombosis; PE=Pulmonary embolism; pre-op=preoperatively; VTE=venous thromboembolism.

<sup>a</sup>This study was underpowered to identify specific risk factors of in-hospital PEs.

**TABLE 4.** Multivariable logistic regression analysis for ICH within 30 days after surgery.

Predictor	Definition	ICH		
		OR	95% CI	p
ASA-classification	I-II	Ref.	-	-
	III	1.45	0.74-3.11	0.31
	IV-V	3.23	1.50-7.59	0.004
Hypertension		2.27	1.38-3.75	0.001
Weight loss		4.42	1.48-10.67	0.003
Bleeding disorder		3.13	1.16-7.09	0.01
Preoperative SIRS		2.45	0.98-5.21	0.05
Preoperative sodium	135-145	Ref.	-	-
	<135	2.41	1.29-4.26	0.004
	>145	1.14	0.06-5.59	0.90
Operative time	minutes	1.20 <sup>a</sup>	1.07-1.33	<0.001

Abbreviations: ASA=American Society of Anesthesiologists, CI=confidence interval, ICH=intracranial hemorrhage; SIRS = systematic inflammatory response syndrome.

<sup>a</sup>Inflated β-coefficients to odds ratio per 60 minutes increase

## Discussion

VTE is one of the most common major complications and reasons for readmission, and ICH the most common reason for reoperation among patients undergoing craniotomy for a primary malignant brain tumor. This multicenter study provides novel and useful information regarding the timing of these events and identification of high-risk patients. The increased risk of VTE extends beyond the period of hospitalization, especially for PE, whereas ICH occurs predominantly within the first days after surgery. The VTE risk profile depends on the type of VTE (DVT versus PE events) and the clinical setting (during hospitalization versus after discharge).

The patient population in this study was technically classified as those with a primary malignant brain tumor based on ICD-9 codes. Gliomas represent close to 80% of primary malignant brain tumors,<sup>39</sup> however, there is no standard ICD-9 code specific for glioma. The Central Brain Tumor Registry of the United States (CBTRUS) argues that multiple combinations of ICD histology codes can be used to define gliomas, and their approach was modeled in this study.<sup>39</sup> Therefore, these results are primarily applicable to glioma patients and should be put in the context of previous outcome research on glioma patients.

### Prior work

Rolston et al. confirmed that the prevalence of VTE following a neurosurgical procedure as registered by the NSQIP registry has remained consistent over the last years.<sup>53</sup> This suggests that perioperative management has not improved effectively with regards to preventing VTE, despite the attention it receives in the neurosurgical literature.

Several multicenter studies have previously investigated the short-term incidence of and risk factors for VTE after brain tumor surgery,<sup>3,13,40-46</sup> of which four studies focused on glioma patients.<sup>3,13,40,42</sup> From these studies, the rate of VTE following craniotomy is cited as 3.3-7.5% for glioma patients<sup>3,13,40,42</sup> and 2.3-4.0% for brain tumors patients in general,<sup>41-43,45,47</sup> with a follow-up ranging from solely the initial hospital stay to six weeks after surgery. The 30-day VTE rate was as high as 9.3% when asymptomatic DVTs were included as well.<sup>3</sup> These results are comparable to the VTE rates found in the current study and suggest a higher rate of VTE in glioma patients postoperatively compared to other brain tumors.

Simanek et al. assessed the cumulative incidence of VTE over time after craniotomy for gliomas, demonstrating a major increase in the number of events in the first three

months after surgery; however, no granular insight into the distribution of events within the first few weeks postoperatively was provided due to a low sample size. Neither did this study stratify for the type of VTE or clinical setting.

Risk factors identified for VTE after craniotomy for gliomas are older age, history of craniotomy, history of VTE, coagulopathy, seizures, increased stay on the intensive care unit, prolonged hospital stay, residual tumor tissue, and absence of thromboprophylaxis.<sup>3,4,6,12-15</sup> Missios et al. stratified for VTE type demonstrating different risk profiles for postoperative DVT and PE. Male gender and Hispanic ethnicity were predictive for PE, whereas chronic heart failure was predictive for DVT.<sup>3</sup> Other predictors of postoperative VTE identified in the broader group of brain tumor patients were higher BMI, hypertension, functional dependence, lower Karnofsky Performance Scale (KPS) score, motor deficits, ventilator dependence, steroid usage, preoperative sepsis, longer operative times, and higher World Health Organization (WHO) tumor grade.<sup>41,43,44,47,48</sup>

Prophylactic anticoagulation is a commonly used strategy to prevent VTE but should be carefully balanced against the risk of ICH. In previous studies, the rates of ICH following craniotomy for a brain tumor is cited as 1.0-4.0% with follow-up ranging between the initial hospital stay and long-term survival after surgery.<sup>6,13-15,45,48-50</sup> However, definitions for major ICH varied based on volumetric measurements of the hematoma, presence of symptoms, decrease in hemoglobin, or need for surgical evacuation.<sup>14,15,22,24,25,51</sup>

Mantia et al. assessed the cumulative incidence of ICH over time after craniotomy for glioma. However, no time-to-event analysis was provided for the direct postoperative period due to a low sample size.<sup>24</sup> Neither did this study stratify for the clinical setting of the patient. Risk factors associated with ICH were history of craniotomy, use of bevacizumab, and therapeutic anticoagulation for VTE.<sup>13,21-25</sup> The association between thromboprophylactic anticoagulation and ICH remains to be elucidated.<sup>15</sup>

To our knowledge, the current study is the first large multicenter assessment including a descriptive time-to-event analysis for both VTE and ICH within 30 days after craniotomy for a primary malignant brain tumor. Additionally, it is the first study that uses the NSQIP database to identify predictors of ICH after brain tumor resection. By addressing thrombotic outcomes as well as hemorrhagic outcomes, this study provides a meaningful direction for future research on thromboprophylactic treatment strategies. Lastly, the large sample size allows a stratification of both the descriptive and inferential analyses, demonstrating differences in risk profile and incidence over time based on VTE type and clinical setting.

## **Limitations**

Complication rates found in the current study can be conservative estimates if events were not reported back to the hospitals. VTEs were only coded as events if they were diagnosed and treated, thereby missing asymptomatic and undetected VTEs. Tumor specific information (histology, size, location, residual tumor volume) and complication specific information (location and classification of DVT, PE, and ICH) was not available. However, both VTE and ICH were defined in the NSQIP database as complications requiring medical and surgical treatment, respectively, resulting in selection of the most clinically relevant events. Perhaps most importantly, no data is available regarding the use of thromboprophylaxis and non-pharmaceutical prophylactic strategies. Therefore, this study offers limited insight into the efficacy of different thromboprophylactic treatment strategies and their association with the occurrence of ICH. Selection bias can be introduced since institutions can selectively contribute patients to the NSQIP registry. Stratifying the analyses based on VTE type and clinical setting reduced the number of events per outcome measure. Yet, our study was not underpowered for most outcome measures according to rule of ten events per variable in the multivariable analysis.<sup>52</sup> Lastly, VTE and ICH events after the 30-day time period established in NSQIP are not accounted for in this study, whereas studies have demonstrated that the risk of VTE events remains non-negligible beyond 30 days postoperatively with incidences up to 26% in the first 12 months postoperatively.<sup>4-6</sup> Despite these limitations, this study provides useful insight into the rates, timing, and predictors of DVT, PE, and ICH after craniotomy for a primary malignant brain tumor. Due to the multicenter nature of the NSQIP dataset, the results of this study may be more representative of typical management at all hospitals, including but not limited to tertiary care academic centers.

## **Implications**

The significant prevalence of VTE and ICH following craniotomy for primary malignant brain tumors found in the current study indicates that there is still room for improvement when it comes to monitoring and preventing these events.

These results particularly encourage the need for continued awareness for VTE post-discharge, and PE in particular. These PEs can also be considered more sudden since they were less often preceded by a detected DVT. PEs preceded by a DVT, however, suggest inadequate treatment of the initial DVT event. It is possible that PE are less frequently preceded by a DVT after discharge. It is our primary suspicion, however, that DVTs remain undetected more frequently after leaving the hospital because they are less frequently symptomatic and cannot be effectively screened for. It is also possible

that patients who develop symptomatic DVTs are unaware of the signs and symptoms until they progress to PE, implicating a possible role for improved patient education in preventing morbidity caused by DVT and PE. In prospective randomized control trials investigating different VTE prophylaxis modalities, Goldhaber et al. screened all craniotomy patients prior to discharge and found 9.3% of patients to have VTE, most of which were asymptomatic in both studies.<sup>46</sup>

Most guidelines recommend that prophylactic use of low-molecular weight heparin or unfractionated heparin should be considered in all cancer patients undergoing major surgery.<sup>16-19</sup> In patients undergoing operations for brain tumors, however, the benefits of anticoagulation should be carefully balanced against the risk of ICH.<sup>54,55</sup> Although most guidelines support the use of pharmacological prophylaxis in patients with brain tumors, proper timing of prophylaxis remains controversial and the use of anticoagulation often depends on the surgeon's preference.<sup>54-56</sup> Recommendations vary between administration throughout hospitalization,<sup>19</sup> up to five to ten days after surgery,<sup>16,17,57</sup> until the patient is mobile,<sup>54</sup> or timing based on the individual risk profile.<sup>58</sup> A lack of scientific evidence is primarily the cause of this variation in recommendations. Recent systematic reviews and meta-analyses of VTE prophylaxis in patients undergoing craniotomy for a brain tumor have been performed.<sup>59-63</sup> These analyses have compared different VTE prophylaxis modalities, as well as their safety and cost effectiveness, but they do not thoroughly investigate the efficacy over time to determine a recommended time frame for thromboprophylaxis. Only one clinical trial studied the effect of continued prophylaxis up to 12 months after surgery.<sup>15</sup> No significant association was found between prolonged prophylaxis and the rate of both VTE and ICH; however, the trial was stopped early because of expiration of study medication, and the control group received placebo instead of short-term prophylaxis. Many patients may not need or benefit from continuing thromboprophylaxis beyond discharge. Algattas et al. reviewed the safety and effectiveness of several thromboprophylactic strategies and indicated that different regimens may have different efficacies depending on the patient's VTE risk profile.<sup>59</sup> This highlights the importance of using the appropriate risk profile for optimizing postoperative management.

Since the NSQIP data does not contain information on thromboprophylactic strategies, the current study provides limited insight into the efficacy or safety of prophylactic anticoagulation and insufficient evidence to change the current clinical practice with regards to thromboprophylaxis. Therefore, we concur with the current guidelines that recommend pharmaceutic prophylaxis (low-molecular weight heparin or unfractionated heparin) in combination with mechanical prophylaxis (anti-embolism

stockings or intermittent pneumatic compression devices) postoperatively until the end of hospitalization or until the patient is mobile. Absolute contra-indications for these include recent ICH or another active major bleeding.<sup>16-19,54,55,57,58</sup>

## **Future research**

Despite its limitations, this study provides useful insight into the prevalence, timing, and risk factors of postoperative VTE and ICH after craniotomy for a primary malignant brain tumor. The results of the current study demonstrate that there is still room for improvement, especially with regards to the prevention of PE after hospitalization. Because of the distinct critical time periods for both thrombotic and hemorrhagic events, it could be worth to investigate the safety and efficacy of continuing prophylactic anticoagulation beyond discharge in high-risk patients. Additionally, the typical patient at risk for developing a VTE during hospitalization is not the same as the typical patient at risk for developing a VTE post-discharge. This is crucial for tailoring post-discharge management to the risk profile of the individual patient. Therefore, future research should study the effects of timing of thromboprophylactic therapy, screening for asymptomatic events, and patient education on the occurrence of VTE and/or ICH. Additionally, future studies should construct prediction models for VTE and ICH and examine the effectiveness of tailoring postoperative thromboprophylaxis to the individual risk profile of patients undergoing craniotomy for a primary malignant brain tumor.

## **Conclusion**

The increased risk of VTE experienced by patients with a primary malignant brain tumor extends beyond the period of hospitalization, especially for PE, whereas ICH occurs predominantly in the first few days after surgery. The risk profile for VTE depends on the type of VTE and clinical setting. VTE can have fatal consequences if not recognized early, therefore clinicians should have high suspicion during the postoperative period and a low threshold for specific monitoring and prevention.

## References

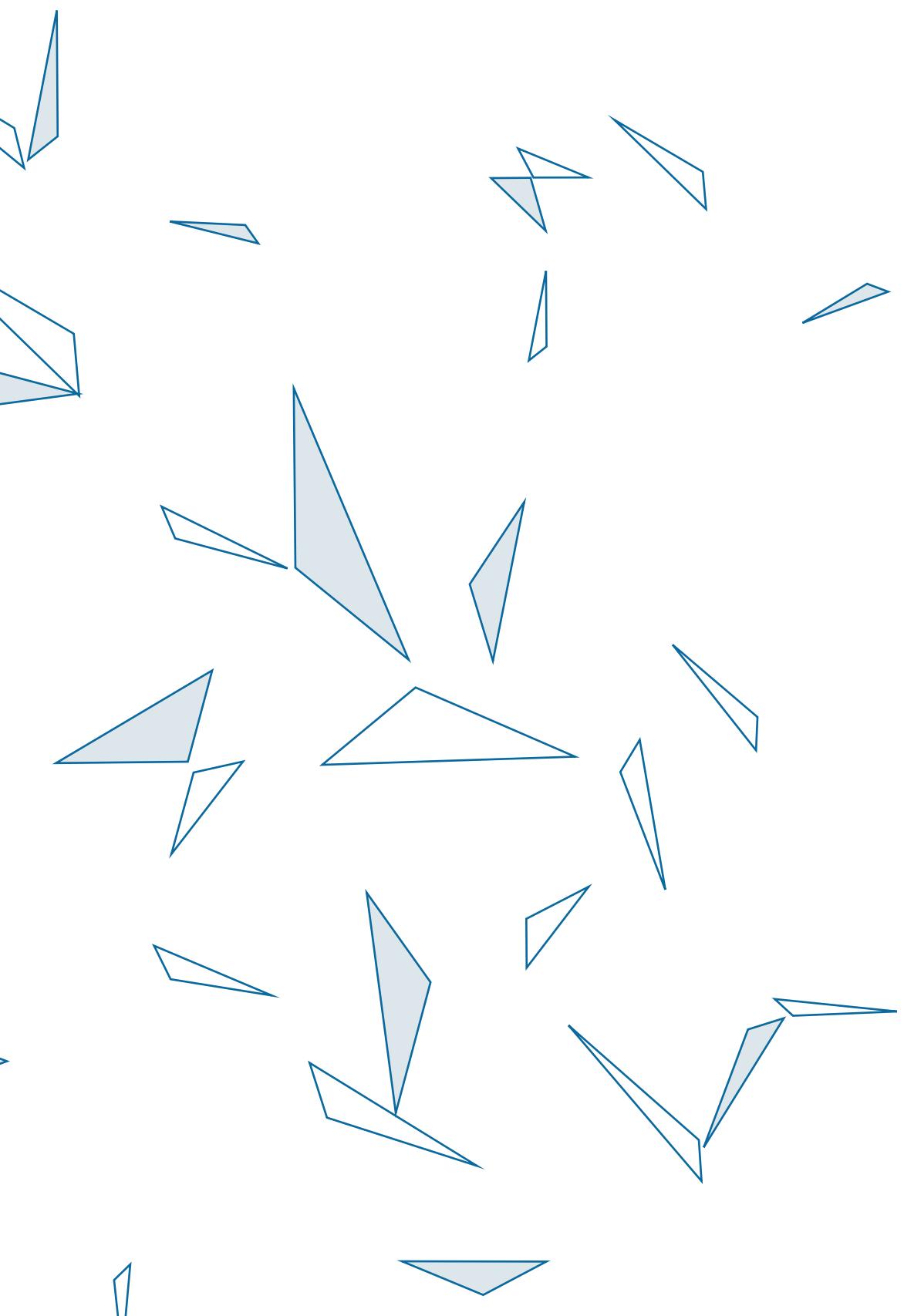
1. Fadul C WJ, Thaler H, Galichich J, Patterson Jr RH, Posner JB. Morbidity and mortality of craniotomy for excision of supratentorial gliomas. *Neurology*. 1988;38:1374-1379.
2. Marcus LP, McCutcheon BA, Noorbakhsh A, et al. Incidence and predictors of 30-day readmission for patients discharged home after craniotomy for malignant supratentorial tumors in California (1995-2010). *Journal of neurosurgery*. 2014;120(5):1201-1211.
3. Missios S KP, Nanda A, Bekelis K. Craniotomy for glioma resection: a predictive model. *World neurosurgery*. 2015;83(6):957-964.
4. Simanek R, Vormittag R, Hassler M, et al. Venous thromboembolism and survival in patients with high-grade glioma. *Neuro Oncol*. 2007;9(2):89-95.
5. Streiff MB, Ye X, Kickler TS, et al. A prospective multicenter study of venous thromboembolism in patients with newly-diagnosed high-grade glioma: hazard rate and risk factors. *Journal of neuro-oncology*. 2015;124(2):299-305.
6. Smith TR, Lall RR, Graham RB, et al. Venous thromboembolism in high grade glioma among surgical patients: results from a single center over a 10 year period. *Journal of neuro-oncology*. 2014;120(2):347-352.
7. Khalil J, Bensaid B, Elkacem H, et al. Venous thromboembolism in cancer patients: an underestimated major health problem. *World J Surg Oncol*. 2015;13:204.
8. Sartori MT, Della Puppa A, Ballin A, et al. Prothrombotic state in glioblastoma multiforme: an evaluation of the procoagulant activity of circulating microparticles. *J Neurooncol*. 2011;104(1):225-231.
9. Cote DJ ST. Venous thromboembolism in brain tumor patients. *Journal of Clinical Neuroscience*. 2016;25:13-18.
10. Stein PD, Beernath A, Meyers FA, Skaf E, Sanchez J, Olson RE. Incidence of venous thromboembolism in patients hospitalized with cancer. *The American journal of medicine*. 2006;119(1):60-68.
11. Kimmell KT WK. Risk factors for venous thromboembolism in patients undergoing craniotomy for neoplastic disease. *Journal of Neuro-oncology*. 2014;120(3):567-573.
12. Edwin NCK, M. N.; Sohal, D.; McCrae, K. R.; Ahluwalia, M. S.; Khorana, A. A. Recurrent venous thromboembolism in glioblastoma. *Thrombosis research*. 2016;137:184-188.
13. Chang SM, Parney IF, McDermott M, et al. Perioperative complications and neurological outcomes of first and second craniotomies among patients enrolled in the Glioma Outcome Project. *Journal of neurosurgery*. 2003;98(6):1175-1181.
14. Perry SL, Bohlin C, Reardon DA, et al. Tinzaparin prophylaxis against venous thromboembolic complications in brain tumor patients. *Journal of neuro-oncology*. 2009;95(1):129-134.
15. Perry JR, Julian JA, Laperriere NJ, et al. PRODIGE: a randomized placebo-controlled trial of dalteparin low-molecular-weight heparin thromboprophylaxis in patients with newly diagnosed malignant glioma. *Journal of thrombosis and haemostasis : JTH*. 2010;8(9):1959-1965.
16. Farge D, Bounameaux H, Brenner B, et al. International clinical practice guidelines including guidance for direct oral anticoagulants in the treatment and prophylaxis of venous thromboembolism in patients with cancer. *Lancet Oncol*. 2016;17(10):e452-e466.
17. Farge D, Debourdeau P, Beckers M, et al. International clinical practice guidelines for the treatment and prophylaxis of venous thromboembolism in patients with cancer. *Journal of thrombosis and haemostasis : JTH*. 2013;11(1):56-70.

18. Lyman GH, Bohlke K, Falanga A, American Society of Clinical O. Venous thromboembolism prophylaxis and treatment in patients with cancer: American Society of Clinical Oncology clinical practice guideline update. *J Oncol Pract.* 2015;11(3):e442-444.
19. Streiff MB. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®). Cancer-Associated Venous Thromboembolic Disease. Version 1.2014. In: National Comprehensive Cancer Network; 2014.
20. Zwicker JI, Karp Leaf R, Carrier M. A meta-analysis of intracranial hemorrhage in patients with brain tumors receiving therapeutic anticoagulation. *Journal of thrombosis and haemostasis : JTH.* 2016;14(9):1736-1740.
21. Zwicker JIKL, R.; Carrier, M. A meta-analysis of intracranial hemorrhage in patients with brain tumors receiving therapeutic anticoagulation. *Journal of thrombosis and haemostasis : JTH.* 2016;14(9):1736-1740.
22. Nghiempuh PLG, R. M.; Pope, W. B.; Lai, A.; Cloughesy, T. F. Safety of anticoagulation use and bevacizumab in patients with glioma. *Neuro-oncology.* 2008;10(3):355-360.
23. Nabi SK, P.; Bozorgnia, F.; Arshad, A.; Mikkelsen, T.; Donthireddy, V. Predictors of Venous Thromboembolism in Patients with Glioblastoma. *Pathology oncology research : POR.* 2016;22(2):311-316.
24. Mantia CU, E. J.; Puligandla, M.; Weber, G. M.; Neuberg, D.; Zwicker, J. I. Predicting the higher rate of intracranial hemorrhage in glioma patients receiving therapeutic enoxaparin. *Blood.* 2017.
25. Al Megren MDW, C.; Al Qahtani, M.; Le Gal, G.; Carrier, M. Management of venous thromboembolism in patients with glioma. *Thrombosis research.* 2017;156:105-108.
26. Lieber BA, Appelboom G, Taylor BE, Malone H, Agarwal N, Connolly ES, Jr. Assessment of the "July Effect": outcomes after early resident transition in adult neurosurgery. *Journal of neurosurgery.* 2015;1-9.
27. McGirt MJ, Godil SS, Asher AL, Parker SL, Devin CJ. Quality analysis of anterior cervical discectomy and fusion in the outpatient versus inpatient setting: analysis of 7288 patients from the NSQIP database. *Neurosurgical focus.* 2015;39(6):E9.
28. Lieber BA, Appelboom G, Taylor BE, et al. Preoperative chemotherapy and corticosteroids: independent predictors of cranial surgical-site infections. *Journal of neurosurgery.* 2015;1-9.
29. Lim S, Parsa AT, Kim BD, Rosenow JM, Kim JY. Impact of resident involvement in neurosurgery: an analysis of 8748 patients from the 2011 American College of Surgeons National Surgical Quality Improvement Program database. *Journal of neurosurgery.* 2015;122(4):962-970.
30. McCutcheon BA, Ciacci JD, Marcus LP, et al. Thirty-Day Perioperative Outcomes in Spinal Fusion by Specialty Within the NSQIP Database. *Spine.* 2015;40(14):1122-1131.
31. Dasenbrock HH, Devine CA, Liu KX, et al. Thrombocytopenia and craniotomy for tumor: A National Surgical Quality Improvement Program analysis. *Cancer.* 2016.
32. Lieber BA, Han J, Appelboom G, et al. Association of Steroid Use with Deep Venous Thrombosis and Pulmonary Embolism in Neurosurgical Patients: A National Database Analysis. *World neurosurgery.* 2016.
33. Kim BD, Smith TR, Lim S, Cybulski GR, Kim JY. Predictors of unplanned readmission in patients undergoing lumbar decompression: multi-institutional analysis of 7016 patients. *Journal of neurosurgery Spine.* 2014;20(6):606-616.
34. Dasenbrock HH, Liu KX, Devine CA, et al. Length of hospital stay after craniotomy for tumor: a National Surgical Quality Improvement Program analysis. *Neurosurgical focus.* 2015;39(6):E12.
35. Sellers MM, Merkow RP, Halverson A, et al. Validation of new readmission data in the American College of Surgeons National Surgical Quality Improvement Program. *Journal of the American College of Surgeons.* 2013;216(3):420-427.

36. Hackett NJ, De Oliveira GS, Jain UK, Kim JY. ASA class is a reliable independent predictor of medical complications and mortality following surgery. International journal of surgery (London, England). 2015;18:184-190.
37. Abt NB, De la Garza-Ramos R, Olorundare IO, et al. Thirty day postoperative outcomes following anterior lumbar interbody fusion using the national surgical quality improvement program database. Clinical neurology and neurosurgery. 2016;143:126-131.
38. Lukasiewicz AM, Grant RA, Basques BA, Webb ML, Samuel AM, Grauer JN. Patient factors associated with 30-day morbidity, mortality, and length of stay after surgery for subdural hematoma: a study of the American College of Surgeons National Surgical Quality Improvement Program. Journal of neurosurgery. 2016;124(3):760-766.
39. Ostrom QT GH, Xu J, Kromer C, Wolinski Y, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2009-2013. Neuro-oncology. 2016;18:v1-v75.
40. Auguste KIQ-H, A.; Gadkary, C.; Zada, G.; Lamborn, K. R.; Berger, M. S. Incidence of venous thromboembolism in patients undergoing craniotomy and motor mapping for glioma without intraoperative mechanical prophylaxis to the contralateral leg. Journal of neurosurgery. 2003;99(4):680-684.
41. Chaichana KLP, C.; Jackson, C.; Martinez-Gutierrez, J. C.; Diaz-Stransky, A.; Aguayo, J.; Olivi, A.; Weingart, J.; Gallia, G.; Lim, M.; Brem, H.; Quinones-Hinojosa, A. Deep venous thrombosis and pulmonary embolisms in adult patients undergoing craniotomy for brain tumors. Neurological research. 2013;35(2):206-211.
42. Chan ATA, A.; Diran, L. K.; Licholai, G. P.; McLaren Black, P.; Creager, M. A.; Goldhaber, S. Z. Venous thromboembolism occurs frequently in patients undergoing brain tumor surgery despite prophylaxis. Journal of thrombosis and thrombolysis. 1999;8(2):139-142.
43. Cote DJD, H. M.; Karhade, A. V.; Smith, T. R. Venous Thromboembolism in Patients Undergoing Craniotomy for Brain Tumors: A U.S. Nationwide Analysis. Seminars in thrombosis and hemostasis. 2016;42(8):870-876.
44. Dasenbrock HHL, K. X.; Chavakula, V.; Devine, C. A.; Gormley, W. B.; Claus, E. B.; Smith, T. R.; Dunn, I. F. Body habitus, serum albumin, and the outcomes after craniotomy for tumor: a National Surgical Quality Improvement Program analysis. Journal of neurosurgery. 2017;126(3):677-689.
45. Nuno MC, C.; Mukherjee, D.; Ly, D.; Ortega, A.; Black, K. L.; Patil, C. G. Association between in-hospital adverse events and mortality for patients with brain tumors. Journal of neurosurgery. 2015;123(5):1247-1255.
46. Goldhaber SZ, Dunn K, Gerhard-Herman M, Park JK, Black PM. Low rate of venous thromboembolism after craniotomy for brain tumor using multimodality prophylaxis. Chest. 2002;122(6):1933-1937.
47. Kimmell KT, Walter KA. Risk factors for venous thromboembolism in patients undergoing craniotomy for neoplastic disease. Journal of neuro-oncology. 2014;120(3):567-573.
48. Smith TRN, A. D., 3rd; Lall, R. R.; Graham, R. B.; McClendon, J., Jr.; Lall, R. R.; Adel, J. G.; Zakarija, A.; Cote, D. J.; Chandler, J. P. Development of venous thromboembolism (VTE) in patients undergoing surgery for brain tumors: results from a single center over a 10 year period. Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia. 2015;22(3):519-525.
49. Portillo Jdir, I. V.; Font, L.; Braester, A.; Madridano, O.; Peromingo, J. A.; Apollonio, A.; Pagan, B.; Bascunana, J.; Monreal, M. Venous thromboembolism in patients with glioblastoma multiforme: Findings of the RIETE registry. Thrombosis research. 2015;136(6):1199-1203.
50. Pan ET, J. S.; Mitchell, S. B. Retrospective study of venous thromboembolic and intracerebral hemorrhagic events in glioblastoma patients. Anticancer research. 2009;29(10):4309-4313.

51. Auer TAR, M.; Marini, F.; Brockmann, M. A.; Tanyildizi, Y. Ischemic stroke and intracranial hemorrhage in patients with recurrent glioblastoma multiforme, treated with bevacizumab. *Journal of neuro-oncology*. 2017.
52. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379.
53. Rolston JD, Han SJ, Bloch O, Parsa AT. What clinical factors predict the incidence of deep venous thrombosis and pulmonary embolism in neurosurgical patients? *Journal of neurosurgery*. 2014;121(4):908-918.
54. Strangman D. Clinical Practice Guidelines for the management of adult gliomas: astrocytomas and oligodendrogiomas. In: Cancer Council Australia/Australian Cancer Network/Clinical Oncological Society of Australia; 2009.
55. Guyatt GH, Akl EA, Crowther M, et al. Executive summary: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 Suppl):7S-47S.
56. Carman TL, Kanner AA, Barnett GH, Deitcher SR. Prevention of thromboembolism after neurosurgery for brain and spinal tumors. *South Med J*. 2003;96(1):17-22.
57. Lyman GH, Bohlke K, Khorana AA, et al. Venous thromboembolism prophylaxis and treatment in patients with cancer: american society of clinical oncology clinical practice guideline update 2014. *J Clin Oncol*. 2015;33(6):654-656.
58. Prevention and treatment of venous thromboembolism. In: International Consensus Statement 2013 Guidelines According to Scientific Evidence / Cardiovascular Disease Educational and Research Trust (UK) / European Venous Forum / North American Thrombosis Forum International Union of Angiology and Union Internationale du Phlebologie; 2013.
59. Algattas H, Darmania D, DeAndrea-Lazarus I, et al. Systematic Review of Safety and Cost-Effectiveness of Venous Thromboembolism Prophylaxis Strategies in Patients Undergoing Craniotomy for Brain Tumor. *Neurosurgery*. 2017.
60. Alshehri N, Cote DJ, Hulou MM, et al. Venous thromboembolism prophylaxis in brain tumor patients undergoing craniotomy: a meta-analysis. *J Neurooncol*. 2016;130(3):561-570.
61. Cote DJ, Dawood HY, Smith TR. Venous Thromboembolism in Patients with High-Grade Glioma. *Semin Thromb Hemost*. 2016;42(8):877-883.
62. Cote DJ, Dubois HM, Karhade AV, Smith TR. Venous Thromboembolism in Patients Undergoing Craniotomy for Brain Tumors: A U.S. Nationwide Analysis. *Semin Thromb Hemost*. 2016;42(8):870-876.
63. Salmaggi A, Simonetti G, Trevisan E, et al. Perioperative thromboprophylaxis in patients with craniotomy for brain tumours: a systematic review. *J Neurooncol*. 2013;113(2):293-303.





# 4

## **Length of thromboprophylaxis in patients operated for a high-grade glioma: A retrospective study**

---

Joeky T. Senders, Tom J. Snijders, Max van Essen, Gaby M. van Bentum,  
Tatjana Seute, Filip Y. de Vos, Pierre A. Robe, Marike L.D. Broekman

**WORLD NEUROSURG. 2018 JUL;115:E723-E730**

# **Abstract**

## **Introduction**

High-grade gliomas (HGG) are associated with venous thromboembolism (VTE). This study investigates the influence of continuing prophylactic anticoagulation post-discharge on the rate of VTE and intracranial hemorrhage (ICH) in patients operated for a HGG.

## **Methods**

All adult patients who underwent sub- or gross-total resection for a HGG at a single institution were included. Multivariable logistic regression analysis was used to investigate the association between the duration of thromboprophylaxis (dalteparin administered 21 versus 0-7 days postoperatively) and the occurrence of VTE and ICH within 21 or 90 days after surgery, corrected for known risk factors.

## **Results**

We included 301 patients, of whom 166 patients received short, and 135 received prolonged thromboprophylaxis. In the multivariable analysis, prolonged thromboprophylaxis was not significantly associated with the occurrence of VTE within 21 days (3.0% versus 1.2%;  $p = 0.24$ ) or 90 days after surgery (8.9% versus 4.8%;  $p = 0.09$ ); however, prolonged prophylaxis was associated with the occurrence of ICH (5.9% versus 0.6%;  $p = 0.03$ ). Additionally, immobility ( $p = 0.03$ ) and high BMI score ( $p = 0.02$ ) were associated with the occurrence of VTE.

## **Conclusions**

Twenty-one days of prophylactic anticoagulation postoperatively was not associated with a decreased rate of VTE compared to thromboprophylaxis until discharge. ICH was more common with prolonged thromboprophylaxis. These results provide insufficient evidence to extend the duration of prophylaxis beyond hospitalization. Large-scale randomized prospective studies are still needed to clarify the safety, efficacy, and optimal timing of postoperative thromboprophylaxis in HGG patients.

# Introduction

Cancer patients are at increased risk for venous thromboembolism (VTE). This risk is especially high in brain tumor patients.<sup>1</sup> Among the different brain tumors, high-grade gliomas (HGG) seem particularly at risk for developing a VTE,<sup>2-4</sup> with reported incidences of symptomatic VTE up to 37% throughout the course of the disease, depending on the follow-up time and prophylactic treatment given.<sup>4-18</sup> Although it is controversial whether VTE reduces survival in HGG patients,<sup>2,4,6-8,14,19</sup> VTE certainly reduces their quality of life and remains one of the main reasons for readmission within 30 days after surgery.<sup>20</sup>

Besides neurosurgery, many other patient, tumor, and treatment related risk factors for the development of VTE have been identified in HGG patients including old age,<sup>3,5,8,21</sup> male sex,<sup>15</sup> obesity,<sup>15</sup> history of VTE,<sup>6,15</sup> blood group A or AB compared to O,<sup>5</sup> elevated factor VIII,<sup>19</sup> low Karnofsky Performance Scale (KPS) score,<sup>3,15,21</sup> paresis,<sup>2-4,17,18</sup> seizures,<sup>6</sup> glioblastoma histology,<sup>4,8</sup> large tumor size,<sup>5</sup> supra-tentorial location,<sup>2</sup> intra-luminal thrombosis in glioma vessels,<sup>22</sup> craniotomy,<sup>3,7,8,12</sup> initial biopsy before resection,<sup>19</sup> residual tumor tissue after surgery,<sup>7</sup> increased postoperative stay on the ICU<sup>6</sup> or in the hospital,<sup>21</sup> number of hospital admissions,<sup>21</sup> steroid usage,<sup>15</sup> chemotherapy,<sup>18,23</sup> and anti-VEGF therapy.<sup>21</sup>

Most guidelines recommend the use of low-molecular weight heparins (LMWHs), often in combination with compression stocking and/or intermittent pneumatic compression, in patients operated for a brain tumor to reduce the risk of VTE; however, proper timing of prophylaxis remains controversial and varies between administration throughout hospitalization,<sup>24</sup> up to 7-10 days after surgery,<sup>25-27</sup> until the patient is mobile,<sup>28</sup> and timing based on the patient's risk profile or the surgeon's preference.<sup>29,30</sup> A lack of scientific evidence is primarily the cause of this variation in recommendations, and the risk of intracranial hemorrhage (ICH) make clinicians lean towards a more conservative thromboprophylactic strategy.<sup>31</sup> A recent study demonstrated, however, that the risk of VTE remains considerably high after discharge, especially for pulmonary embolism, whereas ICH occurred predominantly during hospitalization.<sup>32</sup> This suggests a potential role for continuing LMWH administration beyond discharge.

In our institution, the duration of postoperative thromboprophylaxis has been prolonged up to 21 days after surgery for an extended period of time. This provides the opportunity to assess the effectiveness of this policy and make a direct comparison with the conventional strategy, prophylactic anticoagulation administered until

discharge from the hospital. In this retrospective cohort study, we assessed whether prolonged thromboprophylaxis decreases the rate of postoperative VTE compared to short prophylaxis and investigated its association with the occurrence of ICH.

## **Methods**

### **Subjects**

All adult patients who were operated for a HGG (WHO grade III or IV) at University Medical Center Utrecht, The Netherlands, between the 1<sup>st</sup> of January 2007 and 30<sup>th</sup> of June 2013 were eligible for this study. Exclusion criteria were: age under 18 years at time of surgery and a previous craniotomy. A medical ethics committee stated that the national laws of the medical research human subjects act did not apply to this study because of the retrospective study design; hence, no written informed consent had to be obtained.

### **Outcomes**

The primary outcomes were the occurrence of VTE (within 21 days and 90 days) as well as ICH postoperatively. VTE was defined as clinical symptoms of deep venous thrombosis or pulmonary embolism confirmed by Doppler ultrasonography or CT angiography, respectively. ICH was defined as a postoperative hemorrhage treated by means of surgical evacuation. Because all ICHs occurred in the direct postoperative period, no distinction was made based on the timing of follow-up.

### **Thromboprophylaxis and other covariates**

Subcutaneous dalteparin (5000 UI per day) was administered according to the surgeon's preference. In 2010, two neurosurgeons started continuing thromboprophylaxis up to 21 days after surgery. The other surgeons kept administering thromboprophylaxis until discharge from the hospital (0-7 days). This difference in policy was driven by the physician's preference rather than by patient characteristics. Other variables collected from the electronic health records were gender, age, BMI, prior history of malignancy or VTE, steroid usage, tumor histology, length of hospitalization, extent of resection reported by the neurosurgeon (sub- or gross-total resection), intraoperative functional mapping, operative time, postoperative Karnofsky Performance Scale (KPS) score, postoperative immobility, and adjuvant chemo- and radiotherapy. Postoperative immobility was defined as weakness in a lower limb and/or walking difficulties.

## Analysis

Univariable analysis was performed to explore the relationship between the independent variables and the occurrence of VTE and ICH. Chemotherapy and radiotherapy were not included in the analysis of VTE within 21 days since these adjuvant therapies were generally started six weeks after surgery. Similarly, KPS score, immobility, and length of stay were not included in the analysis of ICH since these variables were assessed after the occurrence of all ICH events. We considered this to be a result rather than a cause of the occurrence of ICH. Subsequent multivariable analysis was aimed at determining the independent contribution of each variable to the risk of postoperative VTE or ICH.

Chi-square test and independent sample t-test were used in the univariable analysis for categorical and continuous data, respectively. Mann-Whitney U test was used for non-parametric continuous data. Variables associated with VTE in the univariable analysis were included in the multivariable logistic regression analysis, with a liberal threshold ( $p < 0.20$ ) in order to include all determinants with potential value. Firth regression analysis was performed if one of the cells in the multivariable analysis contained zero events. Given our interest in the duration of LMWH thromboprophylaxis, this variable was included in the multivariable analysis automatically. A probability level below 0.05 was considered as statistically significant. The  $\beta$ -coefficients of the continuous variables in the final multivariable models were multiplied to represent the odds ratios and confidence intervals of meaningful and interpretable units for age (per ten years increase) and BMI (per five kg/m<sup>2</sup> increase). All statistical analyses were performed by means of the Statistical Package for Social Science (IBM, version 24) software.

## Results

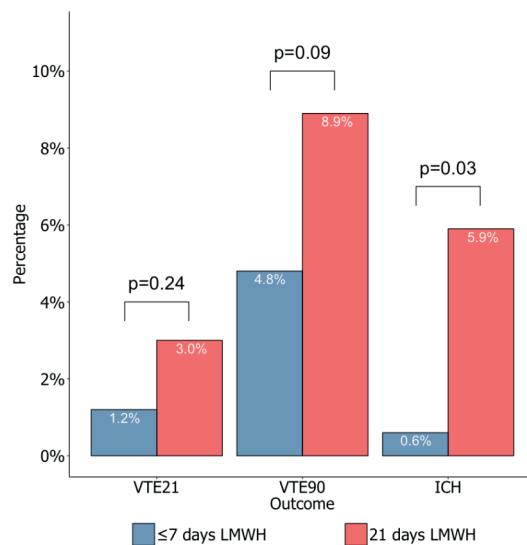
A total of 313 patients underwent craniotomy for a HGG between 2007 and 2013. Twelve patients were excluded because of a previous craniotomy ( $n = 10$ ) or age under 18 years at time of surgery ( $n = 2$ ); therefore, 301 patients were included in the analysis. The cumulative number of patients that developed a VTE within 21 and 90 days were six (2.0%) and 20 (6.6%) patients, respectively. Nine patients (3.0%) had an ICH that required surgical evacuation, all within 10 days after surgery. Hundred-sixty six patients received short prophylaxis, and 135 received prolonged prophylaxis. Baseline characteristics compared by thromboprophylactic regimen are shown in Table 1.

In the univariable analysis for the occurrence of VTE, prolonged compared to short prophylaxis was not significantly associated with the rate of VTE within 21 days (3.0% for the group that received 21 days of dalteparin versus 1.2% for the group that received

0-7 days of dalteparin;  $p = 0.24$ ) (Table 2) or 90 days (8.9% versus 4.8%;  $p = 0.09$ ) (Table 3) (Figure 1). No other risk factors were identified for VTE within 21 days. Patients that developed a VTE within 90 days had a significantly higher BMI compared to patients who did not ( $28.1 \pm 5.8$  versus  $25.9 \pm 4.4 \text{ kg/m}^2$ ;  $p = 0.04$ ).

In the multivariable analysis including LMWH duration and all variables with a p-value less than 0.20, prolonged prophylaxis was not significantly associated with the rate of VTE within 21 days (Odds Ratio[OR] 2.86; 95% CI 0.53 – 21.45;  $p = 0.24$ ) or 90 days (OR 2.19; 95% CI 0.88 – 5.75;  $p = 0.09$ ) after surgery. Additionally, immobility was associated with VTE within 21 days (OR 7.75; 95% CI 1.01 – 43.97;  $p = 0.02$ ) and 90 days after surgery (OR 4.15; 95% CI 1.15 – 13.03;  $p = 0.03$ ). High BMI was associated with VTE within 90 days after surgery (OR 1.66 per 5  $\text{kg/m}^2$  increase; 95% CI 1.08 – 1.83;  $p = 0.02$ ).

In the univariable analysis for ICH, prolonged prophylaxis was significantly associated with the occurrence of ICH (5.9% vs 0.6%;  $p = 0.02$ ) (Table 4). In the multivariable analysis, prolonged prophylaxis was also associated with the occurrence of ICH (OR 9.67; 95% CI 1.73 – 180.95;  $p = 0.03$ ). No other risk factors for ICH were identified.



**FIGURE 1.** Incidence of venous thromboembolism and intracranial hemorrhage compared by duration of postoperative thromboprophylaxis. Abbreviations: ICH=intracranial haemorrhage; LMWH=low-molecular weight heparin; VTE=venous thromboembolism; VTE21=VTE within 21 days after surgery; VTE90=VTE within 90 days after surgery.

**TABLE 1.** Baseline characteristics compared by duration of postoperative thromboprophylaxis.

Patient characteristic	Definition	0-7 days LMWH (n = 166)	21 days LMWH (n = 135)	p
Mean age ± SD	Years	56.5 ± 13.5	59.2 ± 12.7	0.09
Female gender – no. (%)		70 (42.2)	55 (40.7)	0.90
Mean BMI ± SD	Kg/m <sup>2</sup>	26.0 ± 4.8	26.07 ± 4.2	0.89
Histology – no. (%)				
	AA	9 (5.4)	10 (7.4)	0.11
	AOA	4 (2.4)	11 (8.1)	
	AO	2 (1.2)	1 (0.7)	
	GB	151 (91.0)	113 (83.7)	
WHO grade IV – no. (%)		151 (91.0)	113 (83.7)	0.08
IDH1 mutation – no. (%)		8 (7.1)	12 (13.6)	0.19
Median LOS [IQR]	Days	8 [7-10]	7 [6-9]	0.006
Sub-total resection – no. (%)		68 (41.0)	99 (73.3)	<0.001
Awake surgery – no. (%)		5 (3.1)	55 (42.3)	<0.001
Median operative time [IQR]	Min	190 [150-240]	165 [120-210]	0.001
Dexamethasone – no. (%)		148 (89.2)	132 (97.8)	0.007
KPS ≥ 70 – no. (%)		74 (88.1)	97 (82.9)	0.41
Immobility – no. (%)		14 (8.4)	8 (5.9)	0.54
History of VTE – no. (%)		0 (0)	3 (2.2)	0.18
History of malignancy – no. (%)		20 (12.2)	18 (13.3)	0.91
Radiotherapy – no. (%)		149 (91.4)	124 (91.9)	1.00
Chemotherapy – no. (%)		134 (82.2)	100 (74.1)	0.12

Abbreviations: AA=anaplastic astrocytoma; AO=anaplastic oligodendrogloma; AOA=anaplastic oligoastrocytoma; BMI=body mass index; GB=Glioblastoma; GTR=gross-total resection; IQR=inter-quartile range; KPS=Karnofsky performance scale; LMWH=low-molecular weight heparin; LOS=length-of-stay; no.=number; p=p-value; SD=standard deviation; WHO=World Health Organization; VTE=venous thromboembolism

**TABLE 2.** Univariable and multivariable analysis for the outcome of VTE within 21 days.

Univariable Analysis					
Patient characteristic	Definition	Total (n = 301)	No VTE ≤ 21 days (n = 295)	VTE ≤ 21 days (n = 6)	p
Mean age ± SD	Years	57.7 ± 13.2	57.8 ± 13.2	56.8 ± 11.1	0.86
Female gender – no. (%)		125 (41.5)	121 (41.0)	4 (66.7)	0.40
Mean BMI ± SD	Kg/m <sup>2</sup>	26.0 ± 4.5	26.0 ± 4.5	25.5 ± 3.5	0.77
Histology – no. (%)	AA	19 (6.3)	19 (6.4)	0 (0.0)	0.84
	AOA	15 (5.0)	15 (5.1)	0 (0.0)	
	AO	3 (1.0)	3 (1.0)	0 (0.0)	
	GB	264 (87.7)	258 (87.5)	6 (100.0)	
WHO grade IV – no. (%)		264 (87.7)	258 (87.5)	6 (100.0)	0.77
IDH1 mutation – no. (%)		20 (10.0)	20 (10.1)	0 (0.0)	1.00
Median LOS [IQR]	Days	7 [6-10]	7 [6-10]	13 [7-19]	0.37
Sub-total resection – no. (%)		167 (55.5)	164 (55.6)	3 (50.0)	1.00
Awake surgery – no. (%)		60 (20.8)	59 (20.8)	1 (16.7)	1.00
Median operative time [IQR]	Min	180 [130-240]	180 [130-240]	190 [180-230]	0.50
21 days LMWH – no. (%)		135 (44.9)	131 (44.4)	4 (66.7)	0.50
Dexamethasone – no. (%)		280 (93.0)	275 (93.2)	5 (83.3)	0.90
KPS ≥ 70 – no. (%)		171 (85.1)	167 (85.2)	4 (80.0)	1.00
Immobility – no. (%)		22 (7.3)	20 (6.8)	2 (33.3)	0.09
History of VTE – no. (%)		3 (1.0)	3 (1.0)	0 (0.0)	1.00
History of malignancy – no. (%)		38 (12.7)	37 (12.6)	1 (16.7)	1.00

Multivariable Analysis			
Predictor	OR	95% CI	p
Immobility	7.75	1.01 – 43.97	0.02
21 days LMWH	2.86	0.53 – 21.45	0.24

Abbreviations: AA=anaplastic astrocytoma; AO=anaplastic oligodendrogloma; AOA=anaplastic oligoastrocytoma; BMI=body mass index; GB=Glioblastoma; GTR=gross-total resection; IQR=inter-quartile range; KPS=Karnofsky performance scale; LMWH=Low-molecular weight heparin; LOS=length-of-stay; no.=number; OR=odds ratio; SD=standard deviation; WHO=World Health Organization; VTE=venous thromboembolism

**TABLE 3.** Univariable and multivariable analysis for the outcome of VTE within 90 days.

Univariable Analysis					
Patient characteristic	Definition	Total (n = 301)	No VTE ≤ 90 days (n = 281)	VTE ≤ 90 days (n = 20)	p
Mean age ± SD	Years	57.7 ± 13.2	57.6 ± 13.4	59.1 ± 9.4	0.61
Female gender – no. (%)		125 (41.5)	114 (40.6)	11 (5.0)	0.30
Mean BMI ± SD	Kg/m <sup>2</sup>	26.0 ± 4.5	25.9 ± 4.4	28.1 ± 5.8	0.04
Histology – no. (%)	AA	19 (6.3)	19 (6.8)	0 (0.0)	0.39
	AOA	15 (5.0)	15 (5.3)	0 (0.0)	
	AO	3 (1.0)	3 (1.1)	0 (0.0)	
	GB	264 (87.7)	244 (86.8)	20 (100.0)	
WHO grade IV – no. (%)		264 (87.7)	244 (86.8)	20 (100.0)	0.17
IDH1 mutation – no. (%)		20 (10.0)	20 (10.6)	0 (0.0)	0.49
Median LOS [IQR]	Days	7 [6-10]	8 [6-10]	7 [6-8]	0.31
Sub-total resection – no. (%)		167 (55.5)	157 (55.9)	10 (50.0)	0.78
Awake surgery – no. (%)		60 (20.8)	56 (20.7)	4 (21.1)	1.00
Median operative time [IQR]	Min	180 [130-240]	180 [120- 240]	180 [180-200]	0.80
21 days LMWH – no. (%)		135 (44.9)	123 (43.8)	12 (60.0)	0.24
Dexamethasone – no. (%)		280 (93.0)	261 (92.9)	19 (95.0)	1.00
KPS score ≥ 70 – no. (%)		171 (85.1)	158 (84.5)	13 (92.9)	0.65
Immobility – no. (%)		22 (7.3)	18 (6.4)	4 (20.0)	0.07
History of VTE – no. (%)		3 (1.0)	3 (1.1)	0 (0.0)	1.00
History of malignancy – no. (%)		38 (12.7)	36 (12.9)	2 (10.0)	0.98
Radiotherapy – no. (%)		273 (91.6)	254 (91.4)	19 (95.0)	0.88
Chemotherapy – no. (%)		234 (78.5)	216 (77.7)	18 (90.0)	0.31
Multivariable Analysis					
Predictor		OR	95% CI		p
BMI per 5 Kg/m <sup>2</sup> increase		1.66	1.08 – 1.83		0.02
WHO grade IV		6.88	0.87 – 890.69		0.07
Immobility		4.15	1.15 – 13.03		0.03
21 days LMWH		2.19	0.88 – 5.75		0.09

Abbreviations: AA=anaplastic astrocytoma; AO=anaplastic oligodendrogloma; AOA=anaplastic oligoastrocytoma; BMI=body mass index; GB=Glioblastoma; GTR=gross-total resection; IQR=inter-quartile range; KPS=Karnofsky performance scale; LMWH=low-molecular weight heparin; LOS=length-of-stay; no.=number; OR=odds ratio; SD=standard deviation; WHO=World Health Organization; VTE=venous thromboembolism.

**TABLE 4.** Univariable and multivariable analysis for the outcome of postoperative ICH.

Univariable Analysis					
Patient characteristic	Definition	Total (n = 301)	No ICH (n=292)	ICH (n=9)	p
Mean age ± SD	Years	57.7 ± 13.2	57.6 ± 13.1	63.6 (14.0)	0.18
Female gender – no. (%)		125 (41.5)	122 (41.8)	3 (33.3)	0.87
Mean BMI ± SD	Kg/m <sup>2</sup>	26.0 ± 4.5	26.0 ± 4.6	26.2 ± 2.1	0.90
Histology – no. (%)					
	AA	19 (6.3)	19 (6.5)	0 (0.0)	0.73
	AOA	15 (5.0)	15 (5.1)	0 (0.0)	
	AO	3 (1.0)	3 (1.0)	0 (0.0)	
	GBM	264 (87.7)	255 (87.3)	9 (100.0)	
WHO grade IV – no. (%)		264 (87.7)	255 (87.3)	9 (100.0)	0.72
IDH1 mutation – no. (%)		20 (10.0)	20 (10.4)	0 (0.0)	0.53
Sub-total resection – no. (%)		167 (55.5)	161 (55.1)	6 (66.7)	0.73
Awake surgery — no. (%)		60 (20.8)	60 (21.4)	0 (0.0)	0.25
Median operative time [IQR]	Minutes	180 [130-240]	180[133-240]	180 [113-185]	0.36
21 days LMWH – no. (%)		135 (44.9)	127 (43.5)	8 (88.9)	0.02
Dexamethasone – no. (%)		280 (93.0)	271 (92.8)	9 (100.0)	0.87
History of VTE – no. (%)		3 (1.0)	3 (1.0)	0 (0.0)	1.00
History of malignancy – no. (%)		38 (12.7)	37 (12.8)	1 (11.1)	1.00

#### Multivariable analysis

Predictor	OR	95% CI	p
Age per 10 years increase	1.40	0.81 – 2.67	0.26
21 days LMWH	9.67	1.73 – 180.95	0.03

Abbreviations: AA=anaplastic astrocytoma; AO=anaplastic oligodendrogloma; AOA=anaplastic oligoastrocytoma; BMI=body mass index; GB=Glioblastoma; GTR=gross-total resection; IQR=inter-quartile range; KPS=Karnofsky performance scale; LMWH=low-molecular weight heparin; LOS=length-of-stay; no.=number; OR=odds ratio; SD=Standard deviation; WHO=World Health Organization; VTE=venous thromboembolism

## Discussion

Twenty-one days of postoperative LMWH administration was not significantly associated with a lower rate of VTE compared to thromboprophylaxis until discharge (0-7 days); however, prolonged prophylaxis was found to be associated with the occurrence of ICH. Immobility and high BMI were identified as independent predictors of postoperative VTE.

Two case series,<sup>9,10</sup> one retrospective cohort study,<sup>6</sup> and one randomized clinical trial<sup>11</sup> evaluated the effect of prolonged thromboprophylaxis on VTE rate in HGG patients. In the case series and trial performed by Perry et al., thromboprophylaxis was started up to four weeks and continued up to 12 months after surgery.<sup>10,11</sup> The trial closed early because of expiration of study medication and was effectively underpowered to assess the safety and effectiveness of long-term thromboprophylaxis. Additionally, the trial did not directly compare short versus long-term prophylaxis because the control group received placebo instead of short-term prophylaxis. Robins et al. started anticoagulation on the first day of radiotherapy and continued up to 24 months after surgery.<sup>9</sup> Smith et al. administered LMWH during the period of hospitalization postoperatively and prolonged prophylaxis in high-risk patients only.<sup>6</sup> However, the latter study was also underpowered since it only included 25 patients who received prophylactic anticoagulation. Previous studies have already found an association between the therapeutic use of anticoagulation and the incidence of ICH among HGG patients;<sup>31,33,34</sup> however, this has not been demonstrated yet for the prophylactic use of anticoagulation as well as the duration of this treatment. Lastly, immobility<sup>2-4,17,18</sup> and high BMI score<sup>15</sup> have already been identified as predictors of VTE in HGG patients.

Several limitations of the current study should be mentioned. Patients were not randomized to a prophylactic regimen. The duration of thromboprophylaxis was, however, dependent on the timing of surgery and the surgeon's preference. Retrospectively dividing the cohort based on prophylactic regimen can therefore introduce confounding by indication, and this can be a reason for the significant differences in baseline characteristics between the two groups (Table 1). We tried to reduce confounding by including all potential risk factors ( $p<0.20$ ) in the multivariable regression analysis; however, confounders that have not been measured could still influence the results. Due to a low number of events our study can be underpowered, especially for the occurrence of VTE within 21 days after surgery and ICH (all  $< 5$  events per variable in the multivariable analysis). Despite the low number of events, an association with the occurrence of ICH can still be observed, and the absolute VTE rates do not suggest a thromboprophylactic effect of continuing LMWH administration post-discharge. Since all ICHs occurred within 10 days after surgery, it is questionable whether prolongation of thromboprophylaxis is responsible for all these events. Classification into prophylactic subgroups was based on the neurosurgeon performing the operation and the intention-to-treat as described in the postoperative orders; however, some minor degree of discrepancy between these orders and the actual postoperative management cannot be excluded. Lastly, VTEs can be missed if they were asymptomatic or diagnosed in other hospitals and not corresponded back to our

hospital. Given our practical research question, the relevance of asymptomatic cases of VTE is questionable, and missed cases due to loss to follow-up are rare because patients were almost invariably followed in our own hospital.

We think that the limitations are inherently linked to a retrospective study design and proportionate to the strengths of this study. This study focuses on the incidence of VTE in the short-term postoperative period, thereby reducing bias from either adjuvant therapy or a non-representative group of survivors in the long-term period. To our knowledge, this study presents the largest sample of HGG patients prophylactically treated with LMWH after surgery among all studies that address the effect of prolonged thromboprophylaxis in HGG patients. Lastly, the difference in postoperative management between surgeons allows a relatively unbiased comparison between different postoperative strategies based on the duration of thromboprophylaxis.

The results of this study suggest that continuing LMWH beyond hospitalization is neither safe nor effective in preventing VTE. Therefore, we do not recommend prolongation of thromboprophylaxis up to 21 days after surgery routinely in every patient operated for a HGG. However, effectiveness of prolonged thromboprophylaxis targeted to high-risk patients cannot be excluded. Additionally, LMWH administered up to only 21 days can still be too short to achieve significant differences in VTE outcomes measured at 90 days after surgery. Multicenter prospective studies, preferably in a randomized setting, are needed to validate the findings of the current study, and the development of VTE and ICH prediction models can help tailoring postoperative management to the risk profile of the individual patient.

## **Conclusion**

LMWH administration continued up to 21 days after craniotomy for a HGG was not significantly associated with a lower VTE rate compared to prophylaxis until discharge (0-7 days). Prolonged prophylaxis was found to be associated with an increased risk of ICH. Based on our results, we do not recommend prolongation of prophylaxis beyond discharge routinely in every patient operated for a HGG. Future studies are needed to clarify the optimal timing of postoperative thromboprophylaxis and identify HGG patients at risk for VTE and ICH.

## Reference

1. Walker AJ, Card TR, West J, Crooks C, Grainge MJ. Incidence of venous thromboembolism in patients with cancer - a cohort study using linked United Kingdom databases. *Eur J Cancer.* 2013;49(6):1404-1413.
2. Constantini S, Kornowski R, Pomeranz S, Rappaport ZH. Thromboembolic phenomena in neurosurgical patients operated upon for primary and metastatic brain tumors. *Acta neurochirurgica.* 1991;109(3-4):93-97.
3. Chaichana KL, Pendleton C, Jackson C, et al. Deep venous thrombosis and pulmonary embolisms in adult patients undergoing craniotomy for brain tumors. *Neurological research.* 2013;35(2):206-211.
4. Brandes AA, Scelzi E, Salmistraro G, et al. Incidence of risk of thromboembolism during treatment high-grade gliomas: a prospective study. *Eur J Cancer.* 1997;33(10):1592-1596.
5. Streiff MB, Segal J, Grossman SA, Kickler TS, Weir EG. ABO blood group is a potent risk factor for venous thromboembolism in patients with malignant gliomas. *Cancer.* 2004;100(8):1717-1723.
6. Smith TR, Lall RR, Graham RB, et al. Venous thromboembolism in high grade glioma among surgical patients: results from a single center over a 10 year period. *Journal of neuro-oncology.* 2014;120(2):347-352.
7. Simanek R, Vormittag R, Hassler M, et al. Venous thromboembolism and survival in patients with high-grade glioma. *Neuro Oncol.* 2007;9(2):89-95.
8. Semrad TJ, O'Donnell R, Wun T, et al. Epidemiology of venous thromboembolism in 9489 patients with malignant glioma. *Journal of neurosurgery.* 2007;106(4):601-608.
9. Robins HI, O'Neill A, Gilbert M, et al. Effect of dalteparin and radiation on survival and thromboembolic events in glioblastoma multiforme: a phase II ECOG trial. *Cancer Chemother Pharmacol.* 2008;62(2):227-233.
10. Perry SL, Bohlin C, Reardon DA, et al. Tinzaparin prophylaxis against venous thromboembolic complications in brain tumor patients. *Journal of neuro-oncology.* 2009;95(1):129-134.
11. Perry JR, Julian JA, Lapierre NJ, et al. PRODIGE: a randomized placebo-controlled trial of dalteparin low-molecular-weight heparin thromboprophylaxis in patients with newly diagnosed malignant glioma. *Journal of thrombosis and haemostasis : JTH.* 2010;8(9):1959-1965.
12. Pan E, Tsai JS, Mitchell SB. Retrospective study of venous thromboembolic and intracerebral hemorrhagic events in glioblastoma patients. *Anticancer research.* 2009;29(10):4309-4313.
13. Edwin NC, Khouri MN, Sohal D, McCrae KR, Ahluwalia MS, Khorana AA. Recurrent venous thromboembolism in glioblastoma. *Thrombosis research.* 2016;137:184-188.
14. Cheruku R, Tapazoglou E, Ensley J, Kish JA, Cummings GD, al-Sarraf M. The incidence and significance of thromboembolic complications in patients with high-grade gliomas. *Cancer.* 1991;68(12):2621-2624.
15. Yust-Katz S, Mandel JJ, Wu J, et al. Venous thromboembolism (VTE) and glioblastoma. *Journal of neuro-oncology.* 2015;124(1):87-94.
16. Ruff RL, Posner JB. Incidence and treatment of peripheral venous thrombosis in patients with glioma. *Annals of neurology.* 1983;13(3):334-336.
17. Quevedo JF, Buckner JC, Schmidt JL, Dinapoli RP, O'Fallon JR. Thromboembolism in patients with high-grade glioma. *Mayo Clinic proceedings.* 1994;69(4):329-332.
18. Dhami MS, Bona RD, Calogero JA, Hellman RM. Venous thromboembolism and high grade gliomas. *Thrombosis and haemostasis.* 1993;70(3):393-396.
19. Streiff MB, Ye X, Kickler TS, et al. A prospective multicenter study of venous thromboembolism in patients with newly-diagnosed high-grade glioma: hazard rate and risk factors. *Journal of neuro-oncology.* 2015;124(2):299-305.

20. Dickinson H, Carico C, Nuno M, et al. Unplanned readmissions and survival following brain tumor surgery. *Journal of neurosurgery.* 2015;122(1):61-68.
21. Nabi S, Kahlon P, Bozorgnia F, Arshad A, Mikkelsen T, Donthireddy V. Predictors of Venous Thromboembolism in Patients with Glioblastoma. *Pathology oncology research : POR.* 2016;22(2):311-316.
22. Rodas RA, Fenstermaker RA, McKeever PE, et al. Correlation of intraluminal thrombosis in brain tumor vessels with postoperative thrombotic complications: a preliminary report. *Journal of neurosurgery.* 1998;89(2):200-205.
23. Misch M, Czabanka M, Dengler J, et al. D-dimer elevation and paresis predict thromboembolic events during bevacizumab therapy for recurrent malignant glioma. *Anticancer research.* 2013;33(5):2093-2098.
24. Streiff MB. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®). Cancer-Associated Venous Thromboembolic Disease. Version 1.2014. In: National Comprehensive Cancer Network; 2014.
25. Farge D, Bounameaux H, Brenner B, et al. International clinical practice guidelines including guidance for direct oral anticoagulants in the treatment and prophylaxis of venous thromboembolism in patients with cancer. *Lancet Oncol.* 2016;17(10):e452-e466.
26. Lyman GH, Bohlke K, Khorana AA, et al. Venous thromboembolism prophylaxis and treatment in patients with cancer: american society of clinical oncology clinical practice guideline update 2014. *J Clin Oncol.* 2015;33(6):654-656.
27. Farge D, Debourdeau P, Beckers M, et al. International clinical practice guidelines for the treatment and prophylaxis of venous thromboembolism in patients with cancer. *Journal of thrombosis and haemostasis : JTH.* 2013;11(1):56-70.
28. Strangman D. Clinical Practice Guidelines for the management of adult gliomas: astrocytomas and oligodendrogiomas. In: Cancer Council Australia/Australian Cancer Network/Clinical Oncological Society of Australia; 2009.
29. Guyatt GH, Akl EA, Crowther M, et al. Executive summary: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest.* 2012;141(2 Suppl):7S-47S.
30. Carman TL, Kanner AA, Barnett GH, Deitcher SR. Prevention of thromboembolism after neurosurgery for brain and spinal tumors. *South Med J.* 2003;96(1):17-22.
31. Zwicker JI, Karp Leaf R, Carrier M. A meta-analysis of intracranial hemorrhage in patients with brain tumors receiving therapeutic anticoagulation. *Journal of thrombosis and haemostasis : JTH.* 2016;14(9):1736-1740.
32. Senders JT, Goldhaber NH, Cote DJ, et al. Venous thromboembolism and intracranial hemorrhage after craniotomy for primary malignant brain tumors: a National Surgical Quality Improvement Program analysis. *Journal of neuro-oncology.* 2017.
33. Mantia C, Uhlmann EJ, Puligandla M, Weber GM, Neuberg D, Zwicker JI. Predicting the higher rate of intracranial hemorrhage in glioma patients receiving therapeutic enoxaparin. *Blood.* 2017;129(25):3379-3385.
34. Al Megren M, De Wit C, Al Qahtani M, Le Gal G, Carrier M. Management of venous thromboembolism in patients with glioma. *Thrombosis research.* 2017;156:105-108.

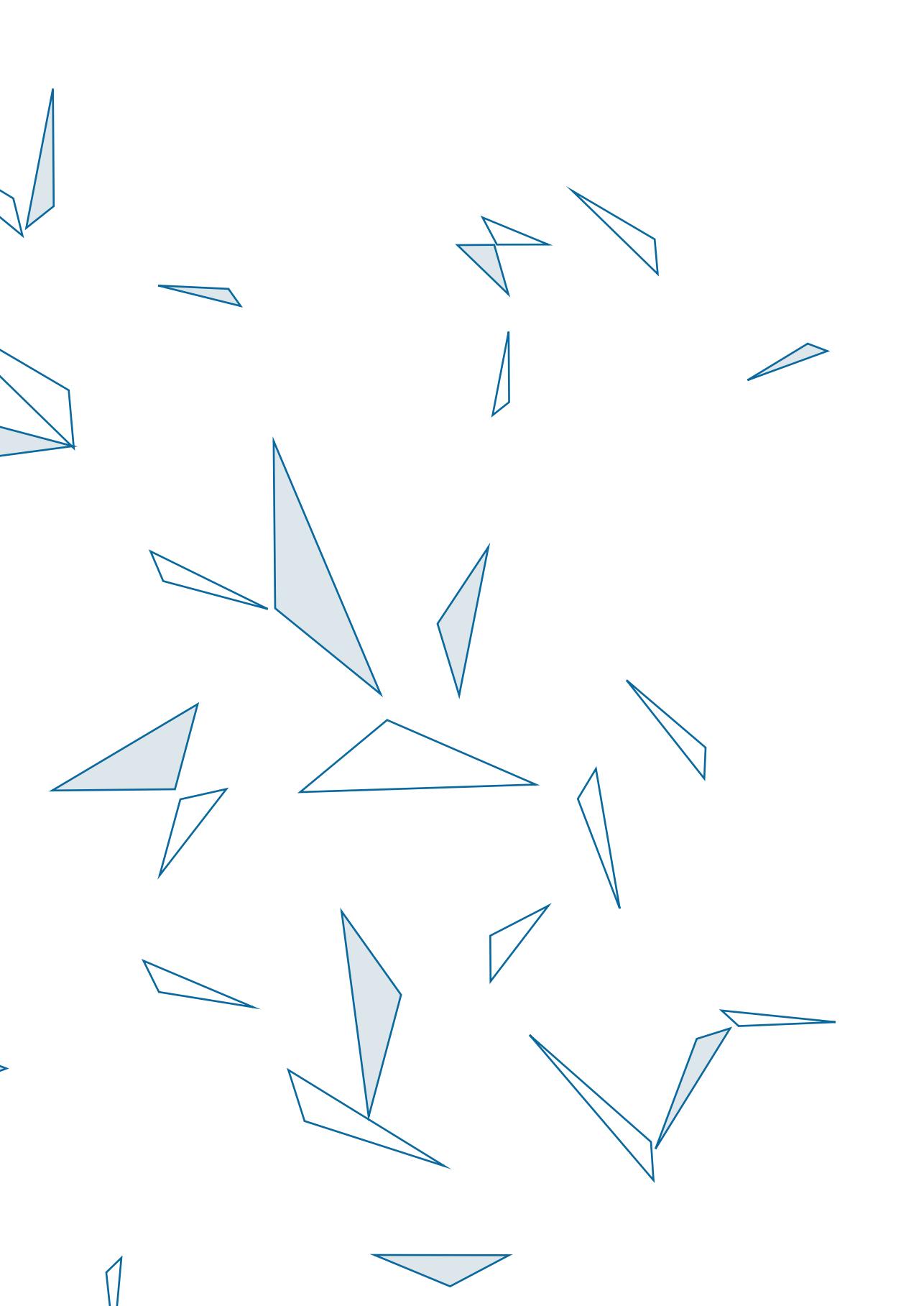


# PART II

---

**Predictive analytics in  
neurosurgical oncology**





# 5

## An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning

---

Joeky T. Senders, Patrick Staples, Alireza Mehrtash, David J. Cote, Martin J.B. Taphoorn, David A. Reardon, William B. Gormley, Timothy R. Smith, Marike L. Broekman\*, Omar Arnaout\*

\*These authors contributed equally and share last authorship

NEUROSURGERY 2020 FEB 1;86(2):E184-E192

# Abstract

## Introduction

Although survival statistics in patients with glioblastoma are well-defined at the group level, predicting individual-patient survival remains challenging due to significant variation within strata. The aim of this study was to compare statistical and machine learning algorithms in their ability to predict survival in glioblastoma patients and deploy the best performing model as an online survival calculator.

## Methods

Patients undergoing an operation for a histopathologically confirmed glioblastoma were extracted from the Surveillance Epidemiology and End Results (SEER) database (2005-2015) and split into a training and hold-out test set in an 80/20 ratio. Fifteen statistical and machine learning algorithms were trained based on 13 demographic, socio-economic, clinical, and radiographic features to predict overall survival, one-year survival status, and compute personalized survival curves.

## Results

In total, 20,821 patients met our inclusion criteria. The accelerated failure time model demonstrated superior performance in terms of discrimination (concordance-index=0.70), calibration, interpretability, predictive applicability, and computational efficiency compared to Cox proportional hazards regression and other machine learning algorithms. This model was deployed through a free, publicly available software interface (<https://cnoc-bwh.shinyapps.io/gbmsurvivalpredictor/>).

## Conclusion

The development and deployment of survival prediction tools require a multimodal assessment rather than a single metric comparison. This study provides a framework for the development of prediction tools in cancer patients, as well as an online survival calculator for patients with glioblastoma. Future efforts should improve the interpretability, predictive applicability, and computational efficiency of existing machine learning algorithms, increase the granularity of population-based registries, and externally validate the proposed prediction tool.

# Introduction

Glioblastoma is the most common primary malignant brain tumor with almost 12,000 new cases per year in the United States and a median survival of only a year after diagnosis.<sup>1</sup> Adequate survival prognostication is essential for informing clinical and personal decision-making. Although survival statistics are well-defined at the group-level, predicting individual patient survival remains challenging due to the heterogenous nature of the disease and significant variation in survival within strata.

In recent years, numerous statistical and machine learning algorithms have emerged that can learn from examples to make patient-level predictions of survival. These algorithms can be particularly useful for tailoring clinical care to the needs of the individual glioblastoma patient.

This study aims to compare the most commonly used statistical and machine learning algorithms in their ability to predict individual-patient survival in glioblastoma patients. In order to promote the reproducibility of the current study and facilitate external validation and implementation of the developed models, we deployed the best performing model as an online calculator that provides interactive, online, and graphical representations of personalized survival estimates.

# Methods

## Data and study population

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement was used for the reporting of this study.<sup>2</sup> Data was extracted from the Surveillance Epidemiology and End Results (SEER) database (2005-2015).<sup>3</sup> The SEER registry compiles cancer incidence and survival data of 18 registries and covers 28% of the U.S. population from academic and nonacademic hospitals, and as such, is broadly representative of the U.S. population as a whole.<sup>4</sup> Patients who underwent surgery for a histopathologically confirmed diagnosis of a glioblastoma (International Classification of Diseases for Oncology-Third Edition [ICD-O-3] codes 9440, 9441, 9442) were included in the analysis. Patients were excluded from the analysis if they died in the direct postoperative period ( $\leq 30$  days after surgery). Our institutional review board has exempted the SEER database from review and waived the need for informed consent due to the retrospective nature of this study.

## Outcome and input features

Although machine learning provides a variety of predictive algorithms, most of them are developed to accommodate binary or continuous outcomes instead of censored survival outcomes (i.e., time-to-event data). To facilitate a vis-à-vis comparison between traditional statistical and novel machine learning algorithms, we compared all algorithms in their ability to predict one or more of the following survival outcomes: (i) continuous: overall survival from diagnosis to death in months, (ii) binary: one-year survival probability, and (iii) censored: subject-level Kaplan-Meier survival curves. All demographic, socio-economic, radiographic, and therapeutic characteristics available at individual patient-level in the SEER registry were included as input features. Continuous variables included age at diagnosis (years) and maximal enhancing tumor diameter in any dimension (millimeters). Categorical variables included sex, race (White, Black, Asian, other), ethnicity (Hispanic, non-Hispanic), marital status (married, non-married), insurance status (insured, uninsured/Medicaid), tumor laterality (left, right, midline), tumor location (frontal, temporal, parietal, occipital, cerebellum, brainstem, ventricles, overlapping lesion), tumor extension (confined to primary location, ventricle involvement, midline crossing), surgery type (biopsy, sub-total resection, gross-total resection), and administration of any form of postoperative chemotherapy and/or radiotherapy. Data on input features and survival outcomes were collected by independent, trained data collectors.

## Statistical analysis

Missing data was multiple imputed by means of a random forest algorithm.<sup>5</sup> The total cohort was randomly split into a training and hold-out test set based on an 80/20 ratio. The Cox proportional hazards regression (CPHR) and the Accelerated Failure Time (AFT) algorithms allow for inferential analysis on censored survival data. Therefore, both approaches were also utilized to provide insight into the independent association between covariates and survival. Interactions between age, sex, surgery type, radiotherapy, and chemotherapy were modeled in both approaches. The Benjamini-Hochberg procedure based on 41 comparisons (26 parameters plus 15 two-way interactions) was used to adjust for multiple testing. The proportional hazards assumption of the CPHR model was assessed by means of the Schoenfeld Residuals Test, and the distribution assumption of the AFT by means of a quantile-quantile plot. All covariates that were statistically significantly associated with survival in the inferential analysis were included in the predictive analysis.

For the predictive analysis, 15 machine learning and statistical algorithms were trained including AFT, bagged decision trees, boosted decision trees, boosted

decision trees survival, CPHR, extreme boosted decision trees, k-nearest neighbors, generalized linear models, lasso and elastic-net regularized generalized linear models, multilayer perceptron, naïve Bayes, random forests, random forest survival, recursive partitioning, and support vector machines.<sup>6-8</sup> Among these, only the AFT, boosted decision trees survival, CPHR, random forest survival, and recursive partitioning algorithms were capable of modeling time-to-event data. Five-fold cross-validation was used on the training set for preprocessing optimization and hyperparameter tuning. Hyperparameters were model-specific, such as the number of trees in a random forest model and the number of layers or nodes per layer in a neural network. The algorithms were subsequently trained with optimized hyperparameter settings on the full training set and evaluated on the hold-out test set, which has not been used for preprocessing and hyperparameter tuning in any form.

## Metrics of predictive performance

Discrimination and calibration were used as metrics for prediction performance. Discrimination reflects the ability of a model to separate observations, whereas calibration measures the agreement between the observed and predicted outcomes.<sup>9</sup> Discrimination was quantified according to the concordance index (C-index). The C-index represents the probability that for any two patients chosen at random, the patient who had the event first is rated as being more at risk of the event according to the model. Therefore, the C-index takes into account the occurrence of the event, as well as the length of follow-up, and is particularly well-suited for right-censored survival analysis.<sup>10</sup> For the subject-level survival curves produced by time-to-event models, the C-index was evaluated per time point weighted according to the survival distribution in the test set and integrated over time. The relationship between predicted one-year survival probability and observed survival rate was graphically assessed in a calibration plot.

## Secondary metrics

In addition to prediction performance, we evaluated additional metrics that pose significant pragmatic challenges to the deployment and implementation of prediction models in clinical care. These metrics include model interpretability, predictive applicability, and computational efficiency. Lack of interpretability is an important concern for the implementation of many machine learning models, which are typically referred to as “black-boxes” and sometimes cited as a weakness compared to classical statistical methods. Inferential utility is a traditional hallmark of model interpretability and therefore included as a model assessment measure. Predictive applicability refers to the type of outcome classes to be predicted (binary, continuous, or time-to-event), as

well as the generated output of the fitted models (class probability, numeric estimate, or subject-level survival curve, respectively). Computational efficiency was measured in terms of model size, loading time, and computation time to produce a prediction. For models that do not provide natural prediction confidence intervals, model predictions were bootstrapped 100 times with replacement to provide such estimates.

We also developed an online, interactive, and graphical tool based on the overall best performing model. Statistical analyses were conducted using R (version 3.5.1, R Core Team, Vienna, Austria).<sup>11</sup> All machine learning modeling was performed using the Caret package,<sup>12</sup> and the application was built and deployed using the Shiny package and server.<sup>13</sup>

## Results

### Patient demographics and clinical characteristics

In total, 20,821 patients met our inclusion criteria. Missing data was multiply imputed for insurance status (16.7% missingness), tumor size (14.3%), tumor laterality (12.0%), tumor location (6.6%), marital status (3.8%), tumor extension (1.6%), surgery type (1.3%), and race (0.2%). Survival time was censored for 3,745 patients (18.0%). The estimated median survival time in the total cohort was 13 months (95%-CI 12-13 months). The total cohort was split into a training and hold-out test set of 16,656 and 4,165 patients, respectively (Table 1).

**TABLE 1.** Baseline characteristics for the training and hold-out test set.

Characteristic	Definition	Training set (n = 16,656)		Hold-out test set (n = 4,165)		p
		n	%	n	%	
Age (years)	<50	2900	17.4	695	16.7	0.505
	50-70	9781	58.7	2456	59.0	
	>70	3975	23.9	1014	24.3	
	Mean ± SD	60.5 ± 13.8		60.7 ± 13.9		
Sex	Female	6872	41.3	1717	41.2	0.982
	Male	9784	58.7	2448	58.8	
Race	White	14821	89.0	3710	89.1	0.509
	Black	1018	6.1	238	5.7	
	Asian	741	4.4	201	4.8	
	Other	76	0.5	16	0.4	
Hispanic	No	14993	90.0	3735	89.7	0.533
	Yes	1663	10.0	430	10.3	
Married	No	5535	33.2	1305	31.3	0.021
	Yes	11121	66.8	2860	68.7	
Insurance	Insured	14503	87.1	3636	87.3	0.717
	Uninsured/Medicaid	2153	12.9	529	12.7	
Laterality	Left	7779	46.7	1901	45.6	0.469
	Right	8714	52.3	2222	53.3	
	Midline	163	1.0	42	1.0	
Location	Frontal lobe	5001	30.0	1219	29.3	0.377
	Temporal lobe	4901	29.4	1270	30.5	
	Parietal lobe	3071	18.4	770	18.5	
	Occipital lobe	875	5.3	206	4.9	
	Ventricle, NOS	79	0.5	14	0.3	
	Cerebellum, NOS	125	0.8	39	0.9	
	Brain stem	75	0.5	12	0.3	
	Overlapping lesion of brain	2529	15.2	635	15.2	
Tumor extension	Confined to primary location	14007	84.1	3536	84.9	0.295
	Ventricles	653	3.9	144	3.5	
	Midline crossing	1996	12.0	485	11.6	
Tumor size (mm)	<25	1539	9.2	382	9.2	0.387
	25-50	9380	56.3	2393	57.5	
	>50	5737	34.4	1390	33.4	

**TABLE 1.** Continued

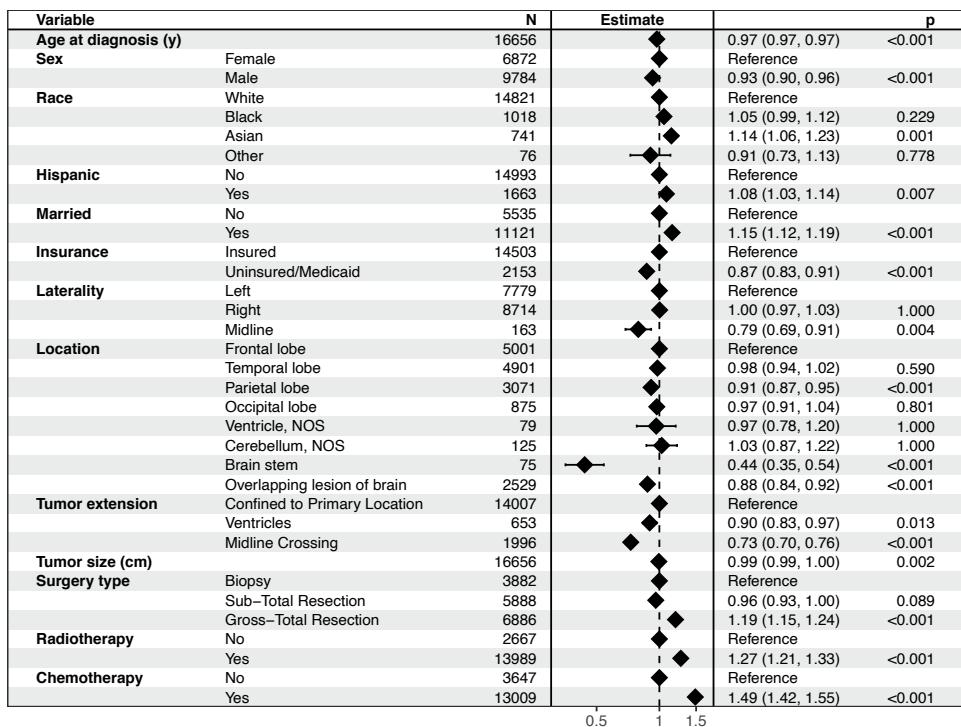
<b>Characteristic</b>	<b>Definition</b>	<b>Training set (n = 16,656)</b>		<b>Hold-out test set (n = 4,165)</b>			<b>p</b>
		<b>n</b>	<b>%</b>	<b>n</b>	<b>%</b>		
Tumor size	Median [IQR]	45 [35-55]	45 [35-55]	45 [35-55]	45 [35-55]		0.986
Surgery type	Biopsy	3882	23.3	977	23.5		0.894
	Sub-total resection	5888	35.4	1456	35.0		
	Gross-total resection	6886	41.3	1732	41.6		
Radiotherapy	No	2667	16.0	662	15.9		0.871
	Yes	13989	84.0	3503	84.1		
Chemotherapy	No	3647	21.9	883	21.2		0.341
	Yes	13009	78.1	3282	78.8		

Abbreviations: IQR=interquartile range; mm=millimeters; n=number; SD=standard deviation

## Inferential analysis

The Schoenfeld residuals test demonstrated that the assumption of proportionality was violated for all variables except sex and ethnicity in the CPHR model (all  $p < .006$  and global test  $p < .001$ ; Supplementary Table S1). The quantile-quantile plot demonstrated a valid log-logistic distribution assumption for the (AFT) model (Supplementary Figure S1). For these reasons, we present the inferential results of the AFT model. The AFT allows for uncomplicated interpretation, as it provides acceleration factors ( $\gamma$ ), which represent the relative survival duration of a strata compared to the reference group. For example, a  $\gamma$  of 1.5 reflects an expected survival duration that is 50% longer compared to the reference group. Multivariable AFT analysis identified older age ( $\gamma = 0.75$  per 10 years increase,  $p < .001$ ), male sex ( $\gamma = 0.93$ ,  $p < .001$ ), uninsured insurance status or insurance by Medicaid ( $\gamma = 0.87$ ,  $p < .001$ ), midline tumors ( $\gamma = 0.79$ ,  $p = .004$ ), tumors primarily located in the parietal lobe ( $\gamma = 0.91$ ,  $p < .001$ ), brain stem ( $\gamma = 0.44$ ,  $p < .001$ ), or multiple lobes ( $\gamma = 0.88$ ,  $p < .001$ ), tumors extending to the ventricles ( $\gamma = 0.90$ ,  $p < .001$ ) or across the midline ( $\gamma = 0.73$ ,  $p < .001$ ), and larger sized tumors ( $\gamma = 0.99$  per cm,  $p < .001$ ) as independent predictors of shorter survival (Figure 1). Asian race ( $\gamma = 1.14$ ,  $p = .001$ ), Hispanic ethnicity ( $\gamma = 1.08$ ,  $p = .007$ ), married marital status ( $\gamma = 1.15$ ,  $p < .001$ ), gross-total resection ( $\gamma = 1.19$ ,  $p < .001$ ), radiotherapy ( $\gamma = 1.27$ ,  $p < .001$ ), and chemotherapy ( $\gamma = 1.49$ ,  $p < .001$ ) were identified as independent predictors of longer survival.

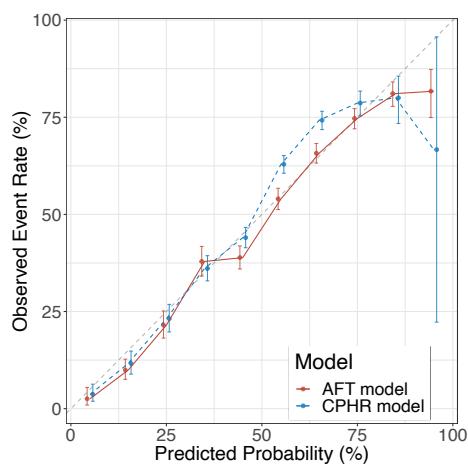
The AFT model with interaction terms demonstrated that age interacted with extent of resection ( $\gamma > 1.03$  per 10 years increase,  $p < .02$ ), as well as radiotherapy ( $\gamma = 1.04$  per 10 years increase,  $p = .03$ ) (Supplementary Table S2).



**FIGURE 1.** Forest plot for the accelerated failure time model characterizing the association between the individual predictors and survival. In the inferential analysis, the estimates for age and tumor size were presented per ten years and ten millimeters increase, respectively, to reflect the incremental relative survival duration of clinically meaningful intervals. The p-value was corrected for multiple testing by means of the Benjamini-Hochberg procedure.

## Predictive analysis

The discriminatory performance on the hold-out test set as measured by the C-index set ranged between 0.66-0.70 and between 0.67-0.70 across all models for predicting overall survival and one-year survival status, respectively (Table 2). Among the time-to-event models, the integrated C-index ranged between 0.68-0.70 for predicting subject-level Kaplan-Meier survival curves. The AFT model based on a log-logistic distribution demonstrated the highest discriminatory performance for computing personalized survival curves. Compared to all continuous and binary models, the AFT model demonstrated similar or better discrimination for predicting overall survival and one-year survival probability, respectively. Model calibration varied significantly across all models (Supplementary Figure S2). The traditional CPHR model systematically underestimated survival in the 0.5-0.75 one-year survival probability range, whereas the AFT model showed better calibration, particularly in this clinically relevant interval (Figure 2).



**FIGURE 2.** Calibration plot demonstrating a systematic underestimation of survival by the Cox proportional hazards regression model in the 0.5 to 0.75 one-year survival probability range and a well-calibrated accelerated failure time model. Abbreviations: AFT=accelerated failure time; CPHR=Cox proportional hazards regression.

**TABLE 2.** Discriminatory performance for all time-to-event, continuous, and binary survival models according to the (integrated) concordance index.

	C-index (95%-CI)		
	Overall survival	1Y-survival status	Integrated C-index
<b>Time-to-event Models</b>			
Accelerated Failure Time	0.70 (0.70-0.70)	0.70 (0.70-0.70)	0.70 (0.70-0.70)
CPHR	0.69 (0.69-0.70)	0.69 (0.69-0.70)	0.69 (0.69-0.70)
Boosted Decision Tree Survival	0.69 (0.69-0.70)	0.69 (0.69-0.70)	0.69 (0.69-0.70)
Random Forest Survival	0.68 (0.68-0.68)	0.69 (0.69-0.69)	0.68 (0.68-0.68)
Recursive Partitioning	0.68 (0.68-0.68)	0.68 (0.68-0.68)	0.68 (0.68-0.68)
<b>Continuous and binary Models</b>			
Gradient Boosting	0.70 (0.70-0.70)	0.70 (0.70-0.70)	NA
Regularized GLM	0.70 (0.70-0.70)	0.70 (0.70-0.70)	NA
GLM	0.70 (0.70-0.70)	0.70 (0.70-0.70)	NA
Support Vector Machines	0.70 (0.70-0.70)	0.69 (0.69-0.69)	NA
Multilayer Perceptron	0.61 (0.61-0.61)	0.69 (0.69-0.69)	NA
Naïve Bayes <sup>a</sup>	NA	0.69 (0.69-0.69)	NA
Random Forest	0.69 (0.69-0.69)	0.69 (0.69-0.69)	NA
Extreme Gradient Boosting	0.68 (0.68-0.68)	0.68 (0.68-0.68)	NA
K-Nearest Neighbors	0.67 (0.67-0.67)	0.68 (0.67-0.68)	NA
Bagging	0.67 (0.66-0.67)	0.66 (0.66-0.66)	NA

Abbreviations: 1Y=one year; C-index=concordance index; CI=confidence interval; CPHR=cox proportional hazards regression; GLM=generalized linear models; NA=not available

<sup>a</sup>Naïve Bayes fits to categorical data only.

## Secondary metrics

Secondary metrics related to model deployment and clinical implementation varied across all models (Table 3). AFT, CPHR, and (regularized) generalized linear models were the only models with inferential utility. AFT, CPHR, boosted decision trees survival, recursive partitioning, and random forest survival were the only models that can analyze time-to-event data and thus compute subject-level survival curves. The application loading time varied between 0.2 seconds and 45 minutes. The 100-fold bootstrapped prediction time varied between 1.9 seconds and four minutes on a single central processing unit.

**TABLE 3.** Secondary metrics for model performance and deployment.

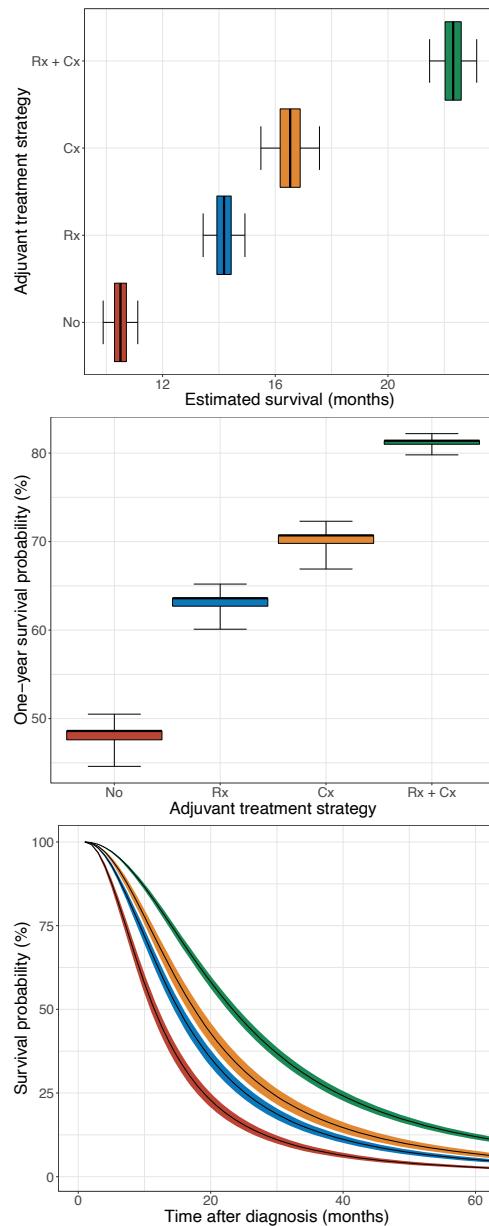
Model	Interpretability		Predictive Applicability			Computational Efficiency <sup>a</sup>		
	Inference	Prediction	Binary	Continuous	Survival Curves	Size (Mb)	Load Time (s)	Prediction Time (s)
AFT	X	X	X	X	X	20	0.9	1.9
Bagging	-	X	X	X	-	16,380	1,335	31.8
Blackboost	-	X	X	X	X	36,790	2,455	234.3
CPHR	X	X	X	X	X	37	1.7	7.5
Recursive Partitioning	-	X	X	X	X	490	52.1	3.4
BDT	-	X	X	X	-	300	8.2	2.1
GLM	X	X	X	X	-	1	0.2	1.7
GLMnet	X	X	X	X	-	109	6.7	2.3
K-Nearest Neighbors	-	X	X	X	-	91	5.6	1.9
Multilayer Perceptron	-	X	X	X	-	45	1.4	17.4
Naïve Bayes	-	X	X	-	-	82	2.9	13.0
Random forest	-	X	X	X	-	1,100	41.4	10.1
Random Forest Survival	-	X	X	X	X	6,350	65.7	139.0
Support Vector Machine	-	X	X	X	-	111	4.8	4.4
XBDT	-	X	X	X	-	92	2.4	1.5

Abbreviations: AFT=accelerated failure time; CPHR=Cox proportional hazards regression; GLM(net)= (Lasso and elastic-net regularized) generalized linear models; Mb = megabyte; s = seconds; TTE=time-to-event; (X)BDT= (extreme) boosted decision trees

<sup>a</sup>Based on a 100-fold bootstrapped model.

## Deployment

Although the AFT model demonstrated similar to superior performance in terms of discrimination and calibration, it outperformed competing statistical and machine learning algorithms in terms of interpretability, predictive applicability, and computational efficiency. Therefore, it was selected as back end for the online survival prediction tool. (<https://cnoc-bwh.shinyapps.io/gbmsurvivalpredictor/>). The estimated survival profile for a hypothetical patient is shown in Figure 3.



**FIGURE 3.** Estimated survival profile of a hypothetical patient (male, 50-years old, white, non-Hispanic, married, insured, left-sided, frontal lobe, confined to its primary location, 50mm in size, gross-total resection), plotted per adjuvant treatment strategy. Personalized estimates of overall survival in months (upper left), one-year survival probability (upper right), and five-year survival curves (lower) as predicted by the accelerated failure time model. The boxes and whiskers in the boxplots represent the 50% and 95% confidence interval, respectively. The ribbons in the survival curves represent the 95% confidence intervals. Abbreviations: Rx=Radiotherapy; Cx=Chemotherapy.

## Discussion

This manuscript and the accompanying online prediction tool provide a framework for individualized survival modeling in patients with glioblastoma that is generalizable to other cancer and neurosurgical patients. Although prior investigation in this area tends to focus on metrics of prediction performance, we advocate a multimodal assessment when constructing and implementing clinical prediction models. The online prediction tool provides interactive, online, and graphical representations of expected survival in glioblastoma patients.

Few other groups have developed an online survival prediction tool for glioblastoma patients.<sup>14–16</sup> Gorlia et al. developed multiple nomograms based on a secondary analysis of trial data using age at diagnosis, World Health Organization performance score (WPS), extent of resection, Mini-Mental State Examination (MMSE) score, and O6-methylguanine–DNA methyltransferase (MGMT) methylation status as input features, thereby achieving a maximum C-index of 0.66.<sup>14</sup> Gittleman et al. developed similar nomograms including sex as an input feature and Karnofsky Performance Status (KPS) score as a measure of functional status. However, model discrimination remained similar (C-index 0.66).<sup>15</sup> Marko et al. developed a model in which extent of resection was modeled as a continuous covariate. This group also utilized an AFT model to account for the violated proportional hazards assumption and achieved a C-index of 0.69.<sup>16</sup> Higher discriminatory performance (C-index 0.63–0.77) was achieved in studies that used machine learning algorithms to analyze complex, high-dimensional data structures, such as genomic, imaging, and health-related quality of life data.<sup>17–25</sup> Although many machine learning algorithms are ideally suited for superior prediction performance by utilizing these high-dimensional data structures, increasing model complexity may incur other costs in terms of interpretability, ease of use, computation speed, and external generalizability.

## Limitations

Due to the retrospective nature of the data acquisition, it cannot be excluded that adjuvant therapy was administered at outside hospitals and not corresponded back to the reporting hospital. However, because of the short survival period in this patient population, the percentage of patients with complete survival follow-up is exceptionally high. Although clinically essential features were included to mitigate the risk of confounding, the possibility of influence from unmeasured confounders cannot be excluded. Randomized data would be ideal; however, it is practically and financially infeasible to establish a cohort on this scale, and it has become ethically unjustifiable

to randomize newly diagnosed patients to a placebo arm now that a proven, effective adjuvant treatment for glioblastoma has emerged.<sup>26</sup> Predictive modeling on this scale remains therefore bound to observational data, thereby highlighting the need for exploring analytical approaches to mitigate confounding.

On average, 3.3% of all data points were missing in the total data set, which was multiply imputed by means of a random forest algorithm to mitigate the risk of systematic bias associated with a complete-case analysis. Nonetheless, survival performance in the current study is limited by the type and number of features included in the SEER registry. As a result, KPS score, isocitrate dehydrogenase 1 (IDH1) mutation, 1p/19q co-deletion, and MGMT methylation status were not included in the current iteration of the prediction model. Despite these limitations, the performance of the current proposed prediction tool exceeds that of the currently available prediction tools and even approximates the performance of many complex radiogenomic models,<sup>17–25</sup> yet with the ease, speed, accessibility, interpretability, and generalizability of clinical prediction tools. Furthermore, this study presents a framework that can be updated and reiterated when novel variables are added to the SEER registry or when novel large-scale multicenter glioblastoma registries are assembled. Because these models are trained on data from thousands of patients from numerous hospitals across the U.S., we expect the fitted models to be less prone to overfitting to data from a single institution and plausibly more generalizable to patients from diverse geographic regions undergoing a variety of clinical treatments.

## Implications

Survival prognostication is critical for clinical and personal decision-making in glioblastoma patients. Although our current prediction tool provides an interactive interface for survival modeling with potential clinical utility, it is designed as a research tool and should not be implemented in clinical practice prior to prospective validation on multiple heterogenous cohorts. Using a population-based registry might be more representative of the typical glioblastoma patient in the US; however, testing the current model on single institutional or multicenter data might be essential to confirm its prognostic value at point-of-care. Furthermore, predictive models should inform rather than direct clinical decision-making. We advocate a multidimensional approach for survival prognostication, in which model predictions are adjusted and balanced against complementary information that is available including clinical experience, neuropsychological testing, imaging data, and genomic information.

Many statistical and machine learning algorithms allow for the analysis of historical patient cohorts to predict survival in new patients. However, prediction performance,

interpretability, clinical utility, computational efficiency, and their associated limitations vary widely across different models due to their mathematical underpinnings. CPHR has emerged as the cornerstone of survival analysis but is limited by the assumption of proportionality, which assumes that the relationship between covariate and outcome is constant over time. In the real world, this association is often dynamic, and the assumption of proportionality is effectively violated. The AFT model does allow for increasing or decreasing covariate risk contribution over time, which is particularly useful in individualizing survival predictions. The AFT model has been shown to be a valuable alternative to CPHR in simulation studies,<sup>27</sup> as well as survival studies on glioblastoma patients.<sup>16</sup>

Molecular markers (e.g., IDH1 mutation, 1p19q codeletion, and MGMT methylation status), as well as functional status (e.g., KPS, MMSE), have been demonstrated to impact survival in glioblastoma patients and are commonly used for stratifying patient cohorts in clinical decision-making. However, they have not yet been included in large-scale, multicenter registries. Inclusion of these variables would improve individual patient survival modeling. Furthermore, granular information with regards to the healthcare setting (e.g., academic versus non-academic) and provided clinical care (e.g., volumetric measurements of tumor size and extent of resection, as well as the timing, type, dose, and sequence of adjuvant treatment) would be valuable to further improve model performance. If addition of any of these variables improves model performance only slightly, however, it may be preferable to exclude some predictors for ease of use at the point of care. Another method to overcome the lack of large-scale granular data sets could be to explore the concept of transfer learning, a common machine learning approach of updating a pre-trained model on novel data sources or even different outcomes.<sup>28</sup> In the context of glioblastoma survival prediction, this could mean developing a base model on population-based data, which is further trained on institutional data to fit institutional patterns and include relevant institutional parameters not available in population-based registries.

Although many machine learning algorithms show great predictive performance, their utility is often limited to continuous and binary models, which merely provide point estimates of overall survival and one-year survival probability at a given point in time, respectively. Transferring the predictive power of these algorithms to time-to-event models allows for the computation of subject-level survival curves, thereby enabling more granular insight into expected survival. Furthermore, time-to-event models can be trained on patients with either complete or incomplete follow-up, which mitigates the systematic bias associated with exclusion of the latter group. Although many machine learning models demonstrate high performance in the

academic realm,<sup>29</sup> lack of interpretability and computational inefficiency hinders their deployment in the clinical realm. When evaluating models for clinical deployment, we recommend evaluating fitted models on several criteria rather than a singular focus on prediction performance since factors unrelated to prediction performance (such as interpretability or applicability) can exclude high-performing models from clinical deployment. Although the AFT model was selected due its high overall performance, the difference in prediction performance was not always clinically meaningful, thereby emphasizing the importance of taking into account these secondary metrics as well. Furthermore, the prediction performance can change as the number and nature of the input features change. For example, the assembly of multimodal data including radiogenomics data might call for alternative analytical approaches in the near future.

Prognostication is and always has been aimed at a moving target and future factors impacting clinical course cannot be modeled, most importantly advances in clinical care. Prediction performance therefore remains an asymptotic ideal for which perfection will never be reached. Future research should focus on developing clinically meaningful and interpretable prediction tools. Improving the end-user transparency regarding the underlying predictive mechanisms and the inherent limitations allows for a safe and reliable implementation of survival prediction tools in clinical care.

## Conclusion

This study provides a framework for the development of survival prediction tools in cancer patients, as well as an online calculator for predicting survival in glioblastoma patients. Future efforts should focus on developing additional algorithms that can train on right-censored survival data, improve the granularity of population-based registries, and externally validate the proposed prediction tool.

## Supplementary material

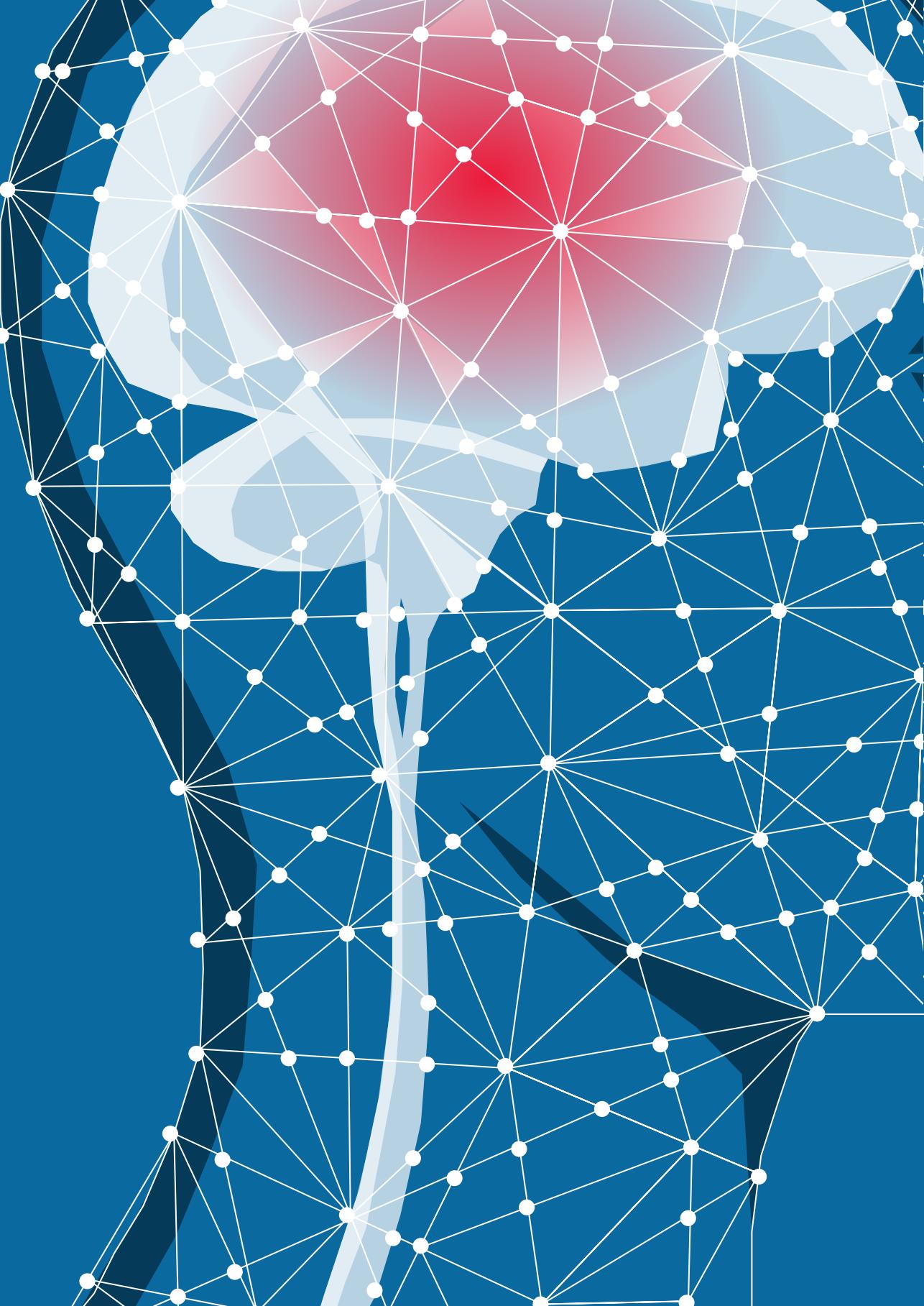
Supplementary tables and figures available online at:

<https://academic.oup.com/neurosurgery/article/86/2/E184/5581744#supplementary-data>

## References

1. Ostrom QT, Gittleman H, Liao P, et al. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro-Oncology*. 2017;19(suppl\_5):v1-v88. doi:10.1093/neuonc/nox158
2. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. doi:10.1136/bmj.g7594
3. Mohanty S, Bilmoria KY. Comparing national cancer registries: The National Cancer Data Base (NCDB) and the Surveillance, Epidemiology, and End Results (SEER) program. *J Surg Oncol*. 2014;109(7):629-630. doi:10.1002/jso.23568
4. Altekruze SF, Rosenfeld GE, Carrick DM, et al. SEER Cancer Registry Biospecimen Research: Yesterday and Tomorrow. *Cancer Epidemiol Biomarkers Prev*. 2014;23(12):2681-2687. doi:10.1158/1055-9965.EPI-14-0490
5. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*. 2013;3(8). doi:10.1136/bmjjopen-2013-002847
6. Senders JT, Arnaout O, Karhade AV, et al. Natural and Artificial Intelligence in Neurosurgery: A Systematic Review. *Neurosurgery*. 2017. doi:10.1093/neuros/nyx384
7. Dietterich TG. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*. Vol 1857. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000:1-15. doi:10.1007/3-540-45014-9\_1
8. Zare A, HOSSEINI M, MAHMOODI M, MOHAMMAD K, ZERAATI H, HOLAKOUIE NAIENI K. A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients. *Iran J Public Health*. 2015;44(8):1095-1102.
9. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
10. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat Med*. 2011;30(10):1105-1117. doi:10.1002/sim.4154
11. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org/>. Published 2008. Accessed June 11, 2018.
12. Kuhn M. Building Predictive Models in R Using the caret Package | Kuhn | Journal of Statistical Software. *Journal of Statistical Software*. 2008. doi:10.18637/jss.v028.i05
13. Chang W, Cheng J, Allaire JJ, et al. Shiny: Web Application Framework for R.; 2018. <https://CRAN.R-project.org/package=shiny>. Accessed June 11, 2018.
14. Gorlia T, Bent MJ van den, Hegi ME, et al. Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. *The Lancet Oncology*. 2008;9(1):29-38. doi:10.1016/S1470-2045(07)70384-4
15. Gittleman H, Lim D, Kattan MW, et al. An independently validated nomogram for individualized estimation of survival among patients with newly diagnosed glioblastoma: NRG Oncology RTOG 0525 and 0825. *Neuro Oncol*. 2017;19(5):669-677. doi:10.1093/neuonc/nox208
16. Marko NF, Weil RJ, Schroeder JL, Lang FF, Suki D, Sawaya RE. Extent of Resection of Glioblastoma Revisited: Personalized Survival Modeling Facilitates More Accurate Survival Prediction and Supports a Maximum-Safe-Resection Approach to Surgery. *JCO*. 2014;32(8):774-782. doi:10.1200/JCO.2013.51.8886

17. Hilario A, Sepulveda JM, Perez-Nuñez A, et al. A Prognostic Model Based on Preoperative MRI Predicts Overall Survival in Patients with Diffuse Gliomas. *American Journal of Neuroradiology*. 2014;35(6):1096-1102. doi:10.3174/ajnr.A3837
18. Cui Y, Ren S, Tha KK, Wu J, Shirato H, Li R. Volume of high-risk intratumoral subregions at multi-parametric MR imaging predicts overall survival and complements molecular analysis of glioblastoma. *Eur Radiol*. 2017;27(9):3583-3592. doi:10.1007/s00330-017-4751-x
19. Mazurowski MA, Desjardins A, Malof JM. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. *Neuro Oncol*. 2013;15(10):1389-1394. doi:10.1093/neuonc/nos335
20. Cui Y, Tha KK, Terasaka S, et al. Prognostic Imaging Biomarkers in Glioblastoma: Development and Independent Validation on the Basis of Multiregion and Quantitative Analysis of MR Images. *Radiology*. 2015;278(2):546-553. doi:10.1148/radiol.2015150358
21. Kickingereder P, Burth S, Wick A, et al. Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. *Radiology*. 2016;280(3):880-889. doi:10.1148/radiol.2016160845
22. Lao J, Chen Y, Li Z-C, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports*. 2017;7(1):10353. doi:10.1038/s41598-017-10649-8
23. Li Q, Bai H, Chen Y, et al. A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme. *Scientific Reports*. 2017;7(1):14331. doi:10.1038/s41598-017-14753-7
24. Mauer MEL, Taphoorn MJB, Bottomley A, et al. Prognostic Value of Health-Related Quality-of-Life Data in Predicting Survival in Patients With Anaplastic Oligodendrogliomas, From a Phase III EORTC Brain Cancer Group Study. *JCO*. 2007;25(36):5731-5737. doi:10.1200/JCO.2007.11.1476
25. Gómez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Palacios-Corona R, Trevino V. Integration and comparison of different genomic data for outcome prediction in cancer. *BioData Mining*. 2015;8(1):32. doi:10.1186/s13040-015-0065-1
26. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine*. 2005;352(10):987-996. doi:10.1056/NEJMoa043330
27. Chiou SH, Kang S, Yan J. Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee. *Journal of Statistical Software*. 2014;61(11). doi:10.18637/jss.v061.i11
28. Pan SJ, Yang Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191
29. Senders JT, Staples PC, Karhade AV, et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurgery*. 2018;109:476-486.e1. doi:10.1016/j.wneu.2017.09.149

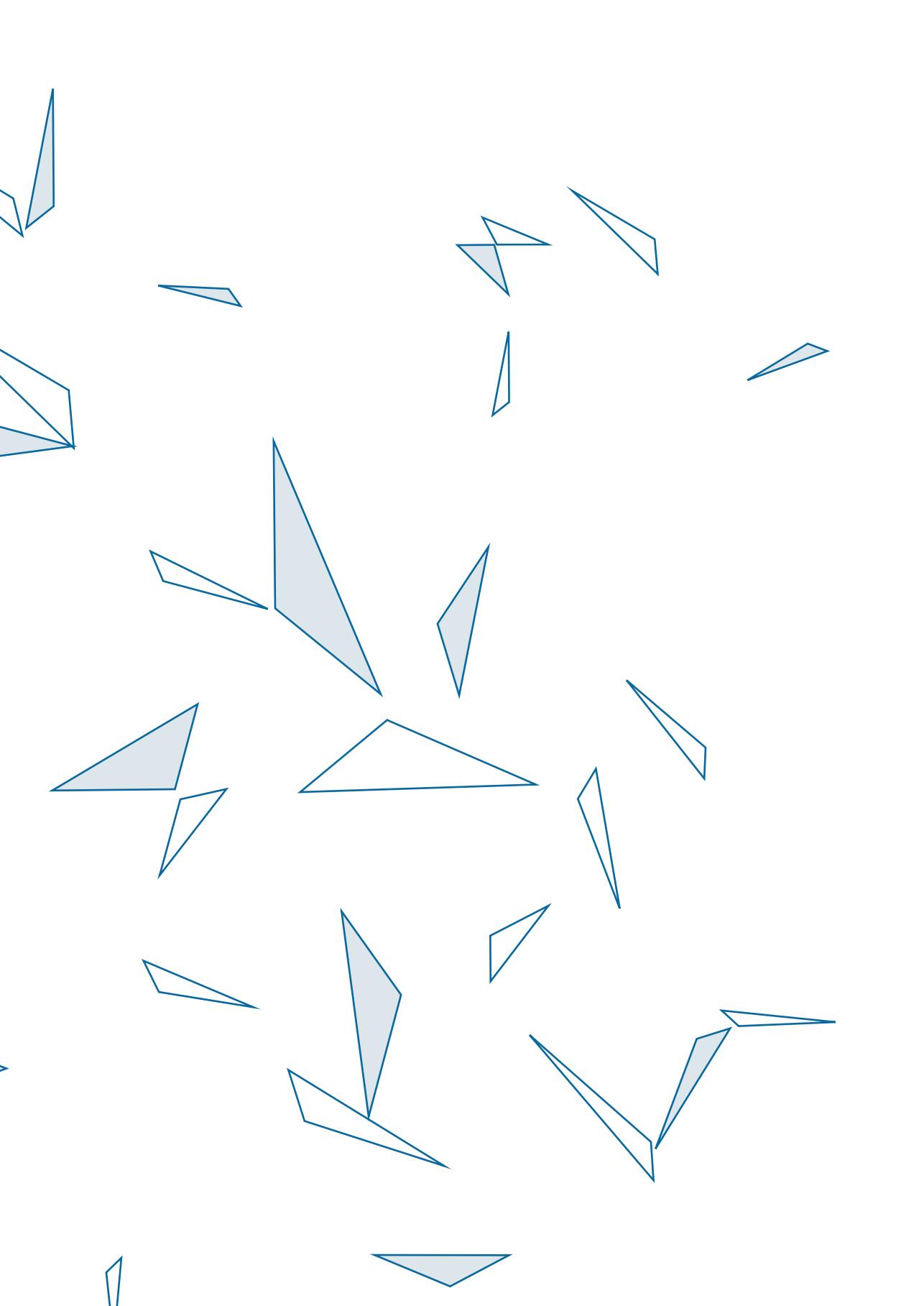




## PART III

---

Natural language processing  
in neurosurgical oncology



# 6

## Automating clinical chart review An open-source natural language processing pipeline developed on free-text radiology reports of glioblastoma patients

---

Joeky T. Senders, Logan D. Cho, Paola Calvachi, John J. McNulty, Joanna L. Ashby, Isabelle S. Schulte, Ahmad Kareem Almekkawi, Alireza Mehrtash, William B. Gormley, Timothy R. Smith, Marike L.D. Broekman, Omar Arnaout

JCO CLINICAL CANCER INFORMATICS 2020 JAN;4:25-34

# **Abstract**

## **Introduction**

The aim of this study was to develop an open-source natural language processing (NLP) pipeline for text mining of medical information from narratively-written reports. Additionally, we aimed to provide insight into the eligibility of variables and the methodological boundaries of text mining in clinical research.

## **Methods**

Various NLP models were developed to extract 15 radiological characteristics from free-text radiology reports of glioblastoma patients. Ten-fold cross-validation was used to optimize the hyperparameter settings and estimate model performance. The Spearman's correlation was calculated to examine how model performance (AUC) was associated with the frequency distribution of the variables of interest and the interrater agreement of the manually provided labels.

## **Results**

In total, 562 unique brain MRI reports were retrieved. NLP extracted 15 radiological characteristics with high to excellent discrimination (AUC 0.82-0.98) and accuracy (78.6-96.6%). Model performance was correlated with the interrater agreement of the manually provided labels ( $\rho=0.904$ ,  $p<0.001$ ) but not with the frequency distribution of the variables of interest ( $\rho=0.179$ ,  $p=0.52$ ). All variables labelled with a near perfect interrater agreement were classified with excellent performance (AUC>0.95). Excellent performance could even be achieved for variables with merely 50-100 observations in the minority group and class imbalances up to a 9:1 ratio.

## **Conclusion**

This study provides an open-source NLP pipeline that allows for text mining of narratively-written clinical reports. Small sample sizes and class imbalance should not be considered as absolute contraindications for text mining in clinical research. However, future studies should report measures of interrater agreement whenever ground truth is based on a consensus label and use this measure to identify clinical variables eligible for text mining.

# Introduction

Analyzing patient characteristics and outcomes can be instrumental for optimizing clinical decision-making. However, most medical information is confined in narratively-written reports, which precludes efficient data gathering and analysis.<sup>1</sup> Manual chart review not only poses substantial costs in terms of time and human resources, variation between and within clinical reviewers can even lead to inconsistencies in data collection and consequently to erroneous inferences from biased study results.<sup>2</sup>

Natural language processing (NLP) provides an automatic and deterministic alternative for the extraction of medical information from free-text clinical reports. It therefore has the potential to accelerate the speed and scale at which clinical research can be performed.<sup>3</sup> Although various pipelines have been developed to automate the extraction of medical information, external validation and optimization of these frameworks is impeded as only a few study groups have released their code on a publicly-accessible repository. Furthermore, the current medical literature on NLP predominantly focusses on the reporting of model performance, whereas it lacks insights into the methodological characteristics that drive model performance and help identify clinical variables eligible for text mining.

In this study, we aimed to develop an open-source NLP pipeline for automated variable extraction using a corpus of free-text radiology reports of glioblastoma patients. Our secondary aim was to provide insight into the feasibility of NLP by studying the statistical properties of the variables of interest. Therefore, we examined how model performance was associated with the frequency distribution of the variables of interest, as well as the interrater agreement of the manually provided consensus labels. These insights can help identify variables eligible, as well as methodological boundaries, for text mining in clinical research.

# Methods

## Retrieval of the free-text radiology reports

This study was conducted and reported according to the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis Or Diagnosis (TRIPOD) statement.<sup>4</sup> The Institutional Review Board of Brigham and Women's Hospital approved the current study and waived the need for informed consent due to its retrospective, observational design. All patients with a histopathologically confirmed diagnosis of a glioblastoma operated at our institution between January 2005 and May 2018 were

included in this study. Glioblastoma constitutes the most prevalent type of primary malignant brain tumor.<sup>5</sup> Patients with glioblastoma generally undergo thorough radiological workup, which is used for diagnostic purposes, as well as neurosurgical planning. The free-text brain MRI reports therefore contain a variety of radiological entities ideally suited to develop an NLP pipeline for clinical text mining. Patients were identified through a departmental database that registers all neurosurgical patients undergoing surgery at our institution. All unique, complete brain MRI reports of the preoperative magnetic resonance imaging (MRI) studies were retrieved by cross-linking the patient identification number with the radiology reports in our centralized institutional data registry. Reports were excluded if the patient underwent any form of oncological treatment (i.e., surgical resection, chemotherapy or radiotherapy) prior to the date of the MRI study or if the reports described lesions suspected for a diagnosis other than a malignant brain tumor.

## Ground truth labels

Ground truth labels of the radiological characteristics of interest were provided manually by clinical reviewers. The total text corpus was divided into two blocks that were each labeled by three independent raters (I.S., J.A., J.M., K.A., L.C., P.C.) for assertions of specific radiological characteristics in a binary fashion (i.e., reported to be present or not). Because each report was labelled by three independent raters, the ground truth was based on the consensus between two or more raters. The radiological characteristics of interest included laterality (left-sided involvement, right-sided involvement, multifocality), location (involvement of the frontal lobe, temporal lobe, parietal lobe, occipital lobe, and corpus callosum), tumor aspect (necrosis, cystic, ring enhancement, heterogenous enhancement), and the presence of other radiological characteristics (hemorrhage, edema, mass effect).

## Preprocessing

Several preprocessing steps were required to convert the brain MRI reports to a numeric format that can be processed by an NLP algorithm. Furthermore, these steps allow for the most parsimonious representation of the lexical content, thereby reducing the feature space and thus the likelihood of overfitting to the training data. Redundant and duplicate text (e.g., date, time, the physician's signature, stop words etc.) was removed, and a Porter stemming algorithm was used to converge words with a similar lexical root.<sup>6</sup> For example, 'necrosis' and 'necrotic' can both be converted to 'necro'. After splitting the stemmed reports into individual words (i.e., tokenization), n-grams were constructed to assign unique value and meaning to adjacent combinations of words.<sup>7</sup> For example, the adjacent words 'ring' and 'enhancement' can be combined

into the 2-gram ‘ring\_enhancement’. Lastly, all text documents were converted to a numeric format by means of the term frequency-inverse document frequency (TF-IDF) vectorizer.<sup>8</sup> This vectorizer converts each text document into a 1-dimensional array of numbers, each of which represent the relative frequency of specific n-grams in a document compared to their prevalence in the total text corpus.

## Hyperparameter tuning

The preprocessed radiology reports were used as input for the NLP algorithm and the consensus-based ground truth labels as the associated outcomes. A least absolute shrinkage and selection operator (LASSO) regression algorithm was used as final classifier due to its speed and regularizing capacity.<sup>9</sup> Because the preprocessing and classification centers around the relative frequency of word or word combinations rather than the order of these words, this NLP approach is considered a bag-of-words approach. In a recent comparative study, we demonstrated that a bag-of-words approach harnessed with a LASSO-regression algorithm outperforms other competing statistical, classical machine learning, and deep learning approaches in classifying free-text radiology reports.<sup>10</sup> Therefore, this approach was utilized in the current study as well. The use of mono-, bi-, tri-, and tetra-grams, size of the vocabulary in the TF-IDF vectorizer, and l2-regularization were presented to the algorithm as hyperparameters, which were optimized by means of 10-fold cross-validation.

## Model evaluation

The NLP model can compute predicted probabilities (number between 0 and 1) representing the probability of an observation belonging to a certain class or directly compute the predicted class in a binary fashion (0 or 1). Predicted probabilities can be used to measure model performance according to the area under the receiver operating characteristic curve (AUC).<sup>11</sup> The AUC is a measure of discrimination and equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example (i.e., the predicted probabilities are higher for reports with the clinical assertions of interest versus reports without). The predicted class can be used to calculate classification accuracy, the percentage of reports classified correctly when the output of the model is binary. Final model performance was calculated as the pooled mean performance and standard deviation of performance across all validation folds.

## Feasibility analysis

To provide insight into the feasibility of text mining for various clinical characteristics, we calculated the correlation between model performance according to the AUC and the statistical properties of the variables of interest (i.e., frequency distribution and interrater agreement of the consensus label). Frequency distribution represents the percentage of observations in the least prevalent outcome group. For example, if a variable is present in 70% of the total cohort, the frequency distribution is represented by the minority group, 30%. As such, a frequency distribution of 50% reflects an equal distribution of observed values, whereas a distribution close to 0% reflects an unequal distribution. Interrater agreement was measured according to the Fleiss' Kappa statistic, which is an extension of the Cohen's Kappa statistic for more than two raters.<sup>12</sup> The Kappa statistic ( $\kappa$ ) accounts for the possibility of agreement occurring by chance and is measured on a scale from -1 to 1. The interpretation of the  $\kappa$  can be categorized according to this scale as less than chance (<0), slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and near perfect (0.81-1).<sup>12</sup> The association between model performance and the statistical properties of the variables of interest was measured according to the Spearman's correlation.

The NLP models were developed and evaluated in Python version 3.6 (Python Software Foundation, <http://www.python.org>) using the Scikit-learn library. The feasibility analysis was performed in R version 3.5.1 (R Core Team, Vienna, Austria, <https://cran.r-project.org>). To promote the transparency and reproducibility of our work, we have released the source code with an open-source license on a publicly-accessible GitHub repository ([https://github.com/jtsenders/nlp\\_glioblastoma](https://github.com/jtsenders/nlp_glioblastoma)). Additionally, a step-by-step pseudocode is provided in Table 1, which can be used to develop similar NLP models for other clinical text mining applications.

## Results

In total, we retrieved 562 unique brain MRI reports of glioblastoma patients operated at our institution. Prevalence of the radiological characteristics reported in the free-text radiology reports ranged between 10.5% for tumor extension into the corpus callosum and 53.7% for left-sided tumor involvement (Table 2).

**TABLE 1.** Pseudocode utilized in the current study in a format that is generalizable to other NLP applications in clinical research.

Phase	Steps
Phase 1: Data Import and preprocessing	A. Import dataframe with the report ID, original report, and binary labels per outcome of interest. B. Randomly shuffle all observations C. In the original report column, subsequently <ul style="list-style-type: none"> <li>a. remove all redundant information (date, time, physician's signature, white spaces between sections, and punctuation between letters) and transform all letters to lower case letters.</li> <li>b. remove all English stop words except 'no' and 'not'</li> <li>c. apply Porter stemmer algorithm</li> </ul> D. Tokenize all reports
Phase 2: Hyperparameter tuning	A. Load preprocessed reports B. Construct hyperparameter grid including the following hyperparameters for <ul style="list-style-type: none"> <li>a. TFIDF vectorization:               <ul style="list-style-type: none"> <li>i. maximal number of features</li> <li>ii. N-gram range</li> </ul> </li> <li>b. LASSO regression algorithm               <ul style="list-style-type: none"> <li>i. <math>L_2</math> regularization</li> </ul> </li> </ul> C. For each grid search (i.e., unique hyperparameter setting) subsequently: <ul style="list-style-type: none"> <li>a. apply the TFIDF vectorizer on the total text corpus</li> <li>b. perform k-fold cross-validation</li> <li>c. calculate the mean performance and standard deviation across all folds</li> </ul>
Phase 3: Compute final results	A. For each outcome, extract the optimal hyperparameter settings based on a single or composite performance metric of interest B. Compute final cross-validated results using optimal hyperparameter settings C. Compute cross-validated ROC plots using optimal hyperparameter settings

Abbreviations: LASSO=Least Absolute Shrinkage and Selection Operator; ROC=receiver operating characteristic curve; TF-IDF=term frequency-inverse document frequency

**TABLE 2.** Descriptive table presenting the prevalence of all radiographic characteristics in the total data set (n = 562 brain MRI reports), as well as the associated interrater agreement for the manually provided labels.

Domain	Subdomain	Frequency		
		n	%	$\kappa^*$
Laterality	left-sided involvement	302	53.7	0.868
	right-sided involvement	281	50.0	0.874
	multipilocality	174	31.0	0.297
Location	frontal lobe	235	41.8	0.847
	temporal lobe	250	44.5	0.831
	parietal lobe	175	31.1	0.813
	occipital lobe	73	13.0	0.821
	corpus callosum	59	10.5	0.574
Tumor aspect	necrosis	165	29.4	0.734
	cystic	85	15.1	0.625
	ring enhancement	122	21.7	0.379
	heterogenous enhancement	232	41.3	0.225
Other characteristic	hemorrhage	151	26.9	0.620
	edema	236	42.0	0.610
	mass effect	288	51.2	0.493

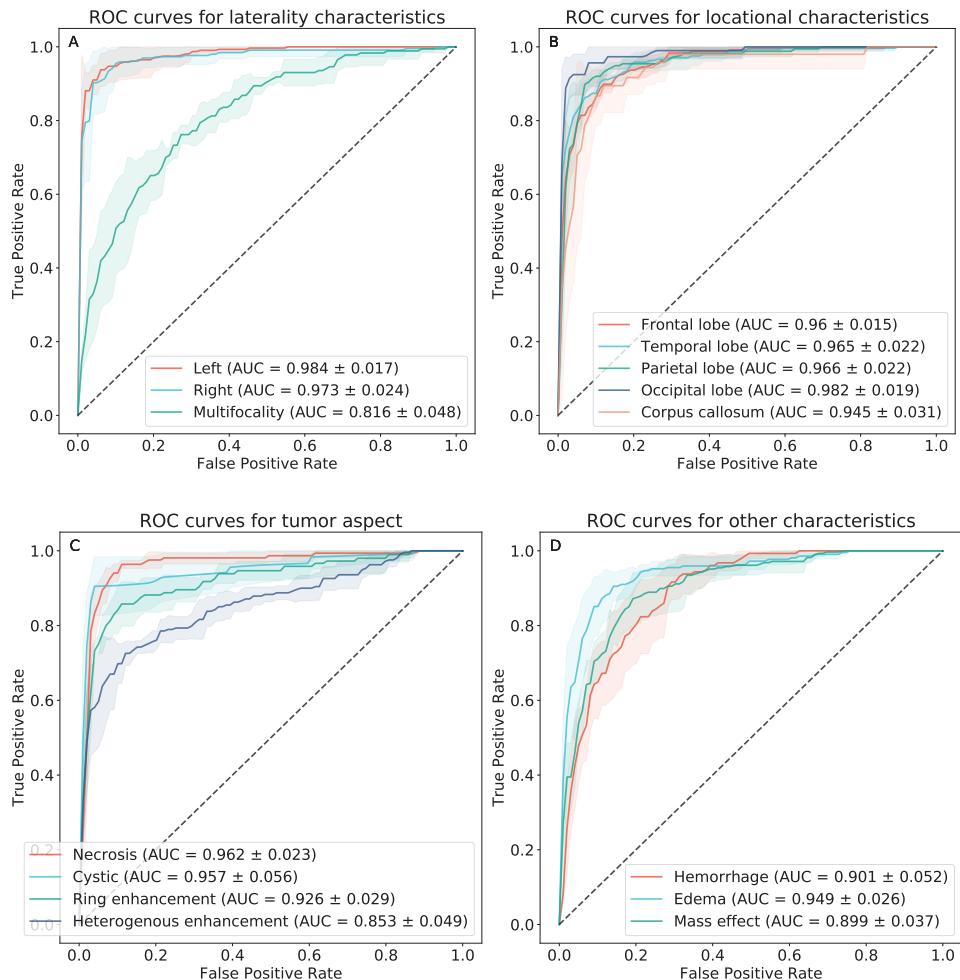
Abbreviations:  $\kappa$ =Fleiss' Kappa statistic

\*The interrater agreement for the consensus labels was calculated by means of the Fleiss' Kappa statistic. The strength of the interrater agreement can be categorized according to this score as less than chance (<0), slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and near perfect (0.81-1).

The overall interrater agreement was substantial ( $\kappa = 0.670$ ) and ranged between fair agreement ( $\kappa = 0.225$  for heterogenous enhancement) to near perfect agreement for ( $\kappa = 0.868$  for right-sided tumor involvement). The cross-validated AUC ranged between 0.816 for multipilocality and 0.984 for left-sided tumor involvement (Table 3, Figure 1), and the binary classification accuracy ranged between 78.6% for multipilocality and 96.6% for tumor involvement of the occipital lobe.

The feasibility analysis revealed that the frequency distribution of the variables of interest was not correlated with model performance ( $\rho = 0.179$ ,  $p = 0.52$ ) (Figure 2a). Excellent model performance (i.e., AUCs > 0.95) could even be achieved for variables with small sample sizes (i.e., as low as 50-100 observations in the minority group) and relatively unbalanced outcomes (i.e., class imbalance up to a 9:1 ratio). In contrast, model performance was strongly correlated with the interrater agreement of the consensus labels ( $\rho = 0.904$ ,  $p < 0.001$ ) (Figure 2b). As the strength of the interrater

agreement increased, model performance according to the AUC increased as well. All six variables that were labelled with a near perfect interrater agreement ( $\kappa$  between 0.8 and 1) were classified with an AUC above 0.95, whereas this performance was achieved for only two out of nine variables with a  $\kappa$  lower than 0.8.



**FIGURE 1.** Receiver operating characteristic curves for all extracted radiological characteristics, grouped by domain including A) tumor laterality B) tumor location C) tumor aspect and D) other characteristics.

**TABLE 3.** Model performance per radiographic feature according to the area under the receiving operating characteristic curve and accuracy.

<b>Domain</b>	<b>Characteristics</b>	<b>AUC (<math>\pm SD</math>)</b>	<b>Accuracy (<math>\pm SD</math>)</b>
Laterality	left-sided involvement	0.984 $\pm$ 0.017	93.6 $\pm$ 3.4
	right-sided involvement	0.973 $\pm$ 0.024	94.1 $\pm$ 3.2
	multifocality	0.816 $\pm$ 0.048	78.6 $\pm$ 4.4
Location	frontal lobe	0.960 $\pm$ 0.015	89.1 $\pm$ 3.5
	temporal lobe	0.965 $\pm$ 0.022	90.9 $\pm$ 3.0
	parietal lobe	0.966 $\pm$ 0.022	91.3 $\pm$ 3.3
	occipital lobe	0.982 $\pm$ 0.019	96.6 $\pm$ 2.2
	corpus callosum	0.945 $\pm$ 0.031	93.1 $\pm$ 3.3
Tumor aspect	necrosis	0.962 $\pm$ 0.023	92.2 $\pm$ 2.0
	cystic	0.956 $\pm$ 0.056	94.0 $\pm$ 2.1
	ring enhancement	0.926 $\pm$ 0.029	89.0 $\pm$ 3.8
	heterogenous enhancement	0.853 $\pm$ 0.049	82.9 $\pm$ 4.8
Other characteristics	hemorrhage	0.901 $\pm$ 0.052	84.0 $\pm$ 6.9
	edema	0.949 $\pm$ 0.026	89.0 $\pm$ 4.1
	mass effect	0.899 $\pm$ 0.037	82.7 $\pm$ 4.8

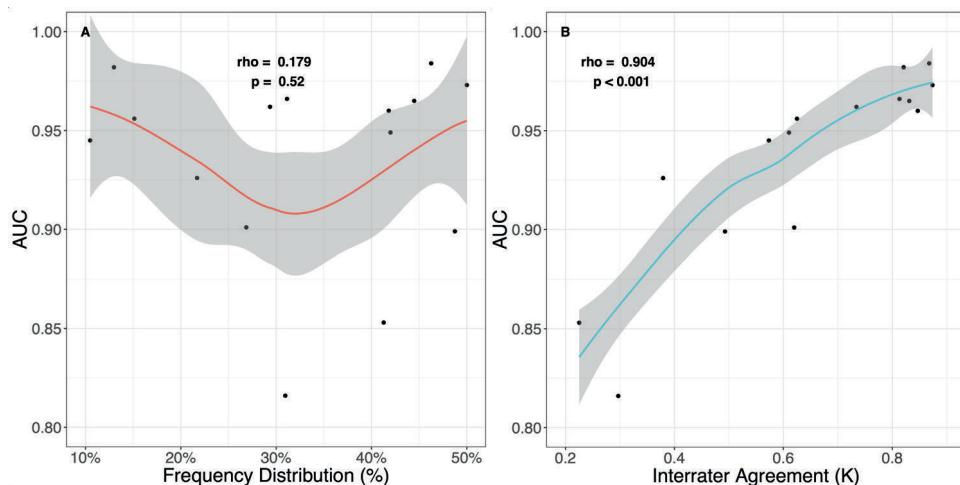
Abbreviations: AUC=area under the receiver operating characteristic curve SD=standard deviation

## Discussion

The aim of this study was to develop an NLP pipeline that allows for automated variable extraction from narratively-written clinical reports. In the current application, NLP was able to extract 15 radiological characteristics from free-text radiology reports of brain MRI studies in glioblastoma patients with high to excellent performance. Model performance was correlated with the interrater agreement of the manually provided labels rather than the frequency distribution of the variables of interest.

Several studies have already developed NLP frameworks for text mining of medical information. Characterizing neoplastic lesions by parsing free-text radiology and pathology reports has been the primary focus in this field of research;<sup>13-18,18-27</sup> however, operative notes, discharge summaries, and outpatient notes are increasingly being analyzed to examine clinical symptoms, postoperative complications, adverse (drug) events, and post-discharge patient follow-up as well.<sup>28-32</sup> Although model performance reported in these studies already exceeds human performance in terms speed and consistency, very few study groups have actually made their code publicly-

accessible.<sup>33-35</sup> This constitutes a significant loss of potential as open-source coding allows for transparency, reproducibility, and external generalizability of the developed NLP pipelines.<sup>36</sup> Furthermore, model performance is often presented as the main finding in the current medical literature, whereas the question why certain variables are more suitable for text mining remains relatively unexposed.



**FIGURE 2.** Scatterplots depicting the correlation between model performance and statistical properties of the variables of interest. Each of the 15 radiological characteristics is represented by a point on these scatter plots. On the y-axis, the performance of the NLP models developed to extract these variables was mapped and measured according to the AUC. On the x-axis, the frequency distribution of the variables (A) and the interrater agreement of the manually provided labels (B) were mapped. The frequency distribution represents the percentage of observations in the least prevalent group, and the interrater agreement was calculated by means of the Fleiss' Kappa statistic ( $\kappa$ ). The association between these statistical properties and model performance was calculated by means of the Spearman's correlation. The smoothed line depicts a Local Polynomial Regression Fitting and the ribbon the associated standard deviation. In the current sample, model performance was statistically significantly correlated with the strength of the interrater agreement of the manually provided labels ( $\rho=0.904$ ,  $p<0.001$ ), but not with the equivalence of the frequency distribution of these variables ( $\rho=0.179$ ,  $p=0.52$ ). Abbreviations: AUC=area under receiver operating characteristics curve;  $\kappa$ =Fleiss' Kappa statistic; % =percentage of patients in the minority group.

## Implications

In the current study, we have developed an open-source NLP pipeline for text mining of medical information using a corpus of free-text radiology reports of patients with a glioblastoma. This pipeline can guide the development of NLP models for other patient cohorts, medical reports, or clinical characteristics as well. Automated extraction of medical information could accelerate the speed and scale at which retrospective chart

review can be done, but also allows for the assembly of large-scale prospective data registries. Both are currently assembled by manual chart review, which is expensive in terms of time and human resources.<sup>2</sup> Significant intra- and interrater variability can be introduced because human coding is subject to fatigue, personal interpretations, biased preconceptions, and progressive insight. Furthermore, inconsistent data collection could even propagate into biased study results and interpretations. The use of consensus labels could attenuate the variation in human coding and develop NLP algorithms that are fast, deterministic, and reproducible by nature.

This study also provides insight into the feasibility of automated extraction of medical information by investigating the correlation between model performance and statistical properties of the variables to extract. These findings suggest that small sample sizes (i.e., as low as 50-100 observations in the minority group) and relatively unbalanced outcomes (i.e., class imbalance up to a 9:1 ratio) should not be considered as limitations or absolute contraindications for NLP modeling. The strong correlation with interrater agreement underlines that a predictive model is as good as the examples it learns from. Interrater agreement might therefore serve as a useful screening tool for the feasibility of text mining on the variables of interest. Furthermore, it might also reflect the lexical complexity of the NLP task at hand. Clinical assertions on left or right-sided tumor involvement might, for example, be less subject to interpretation than higher-level and abstract concepts, such as a patient's perception of quality of life.

## **Limitations**

Several limitations of this study should be mentioned. This pipeline has been developed on a text corpus of radiology reports of a homogenous patient cohort at a single institution, which limits its generalizability to other clinical reports, patient cohorts, and institutions. Instead of the resultant models, the underlying code pipeline was therefore made publicly-accessible in order to promote the reproducibility and external generalizability of the current work. Preserving a residual hold-out test set would have been the most rigorous method for assessing model performance. However, the considerably small text corpus ( $n = 562$  reports) increases the risk of selecting a non-representative sample for final evaluation. To avoid the risk of systematic over- or underestimation of model performance due to a non-representative hold-out test set, model performance was evaluated by means of 10-fold cross-validation which provided a pooled estimate across all validation folds. Although equivalence in the frequency distribution was not significantly associated with model performance in the current feasibility analysis, an association above the examined thresholds (i.e., class imbalance above a 9:1 ratio and less than 50 observations in the minority group) cannot

be excluded. Likewise, other statistical and lexical properties might have a significant influence on model performance and deserve further investigation as well. Despite these limitations, this study provides an open-source NLP pipeline, as well as unique insights into the statistical requirements, for text mining in clinical research.

## Future studies

In order to promote the transparency, reproducibility, and external generalizability of NLP research in healthcare, we advocate that future studies release their source code on a publicly-available repository. Given its strong correlation with model performance, we suggest that future studies report measures of interrater agreement whenever ground truth is based on a consensus rating of human annotators. Further investigating the association between model performance and statistical or lexical properties allows for careful selection of variables eligible for text mining, as well as tailoring the NLP approach to the nature of the clinical text corpus at hand.

## Conclusion

The current study has developed an open-source NLP pipeline for automated variable extraction, which can guide the development of text mining frameworks for other patient cohorts, medical reports, and clinical characteristics as well. In the current sample, model performance was correlated with the interrater agreement of the manually provided labels rather than the frequency distribution of the variables of interest. Class imbalances up to a 9:1 ratio, as well as 50-100 observations in the minority group, should therefore not be considered as contraindications for clinical text mining. Future studies should report measures of interrater agreement whenever ground truth is based on a consensus rating of human annotators and employ open-source coding to promote the transparency, reproducibility, and external generalizability of NLP research in healthcare.

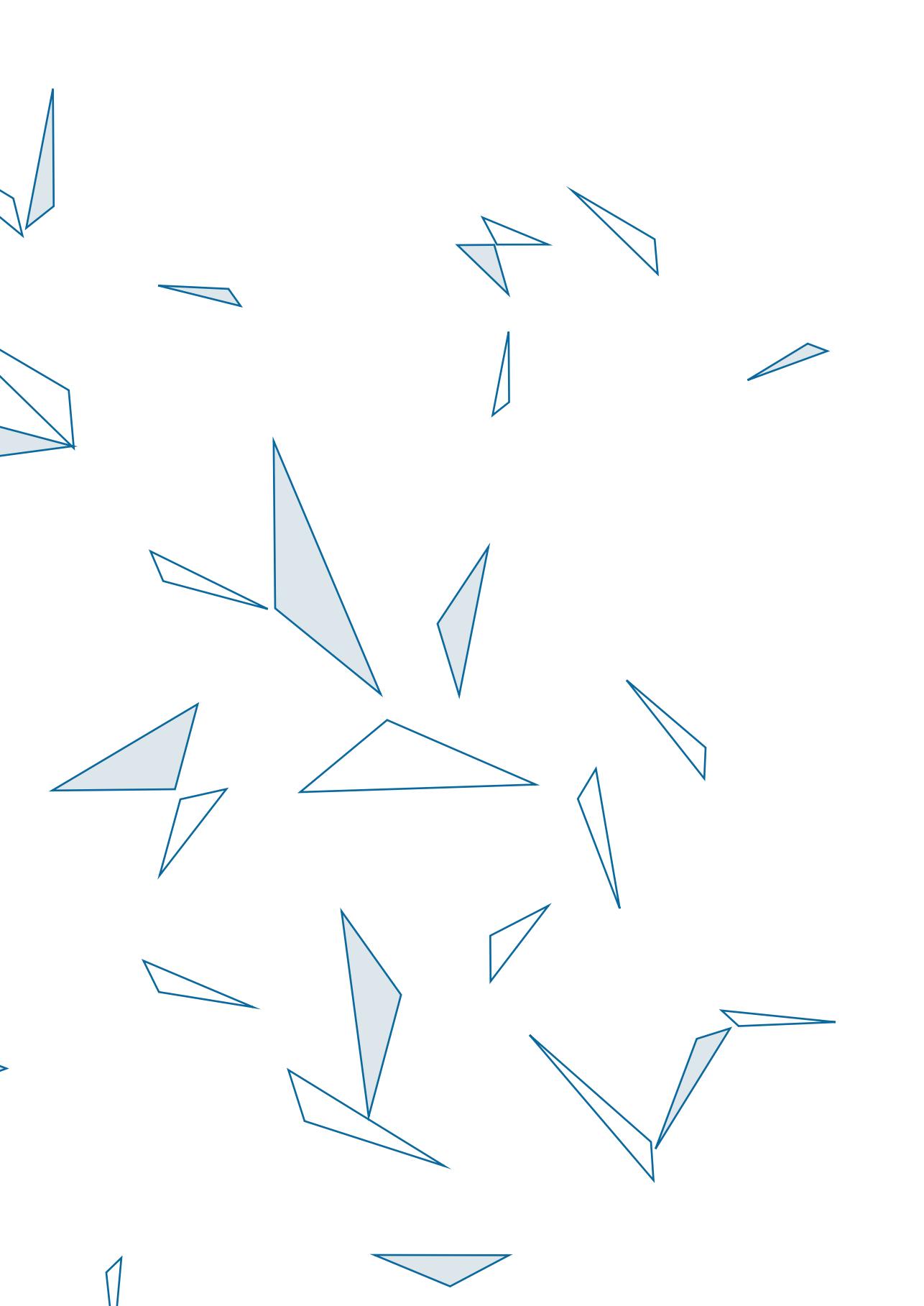
## REFERENCES

1. Ross MK, Wei W, Ohno-Machado L. "Big Data" and the Electronic Health Record. *Yearb Med Inform.* 2014;9(1):97-104. doi:10.15265/IY-2014-0003
2. Matt V, Matthew H. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof.* 2013;10. doi:10.3352/jeehp.2013.10.12
3. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc.* 2011;18(5):544-551. doi:10.1136/ajmajnl-2011-000464
4. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594. doi:10.1136/bmj.g7594
5. Ostrom QT, Gittleman H, Liao P, et al. CBTUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro-Oncology.* 2017;19(suppl\_5):v1-v88. doi:10.1093/neuonc/nox158
6. Porter MF. An algorithm for suffix stripping. *Program.* 2006;40(3):211-218. doi:10.1108/00330330610681286
7. Nguyen VH, Nguyen HT, Duong HN, Snasel V. n-Gram-Based Text Compression. *Comput Intell Neurosci.* 2016;2016. doi:10.1155/2016/9483646
8. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *J Digit Imaging.* 2018;31(2):178-184. doi:10.1007/s10278-017-0027-x
9. Ranstam J, Cook JA. LASSO regression. *BJS.* 2018;105(10):1348-1348. doi:10.1002/bjs.10895
10. Senders JT, Karhade AV, Cote DJ, et al. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. *JCO Clinical Cancer Informatics.* 2019;In Press.
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747
12. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics.* 1977;33(1):159-174. doi:10.2307/2529310
13. Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. *International Journal of Medical Informatics.* 2019;124:6-12. doi:10.1016/j.ijmedinf.2019.01.004
14. Khor RC, Nguyen A, O'Dwyer J, et al. Extracting tumour prognostic factors from a diverse electronic record dataset in genito-urinary oncology. *International Journal of Medical Informatics.* 2019;121:53-57. doi:10.1016/j.ijmedinf.2018.10.008
15. Trivedi HM, Panahiazar M, Liang A, et al. Large Scale Semi-Automated Labeling of Routine Free-Text Clinical Records for Deep Learning. *Journal of Digital Imaging.* 2019;32(1):30-37. doi:10.1007/s10278-018-0105-8
16. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated Extraction of Grade, Stage, and Quality Information From Transurethral Resection of Bladder Tumor Pathology Reports Using Natural Language Processing. *JCO Clinical Cancer Informatics.* 2018;(2):1-8. doi:10.1200/CCI.17.00128
17. Miao S, Xu T, Wu Y, et al. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. *International Journal of Medical Informatics.* 2018;119:17-21. doi:10.1016/j.ijmedinf.2018.08.009
18. Tang R, Ouyang L, Li C, et al. Machine learning to parse breast pathology reports in Chinese. *Breast Cancer Res Treat.* 2018;169(2):243-250. doi:10.1007/s10549-018-4668-3

19. Patel TA, Puppala M, Ogunti RO, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. *Cancer*. 2017;123(1):114-121. doi:10.1002/cncr.30245
20. Pershad Y, Govindan S, Hara AK, et al. Using Naïve Bayesian Analysis to Determine Imaging Characteristics of KRAS Mutations in Metastatic Colon Cancer. *Diagnostics*. 2017;7(3):50. doi:10.3390/diagnostics7030050
21. Schroeck FR, Patterson OV, Alba PR, et al. Development of a Natural Language Processing Engine to Generate Bladder Cancer Pathology Data for Health Services Research. *Urology*. 2017;110:84-91. doi:10.1016/j.urology.2017.07.056
22. Yim W, Denman T, Kwan SW, Yetisgen M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Jt Summits Transl Sci Proc*. 2016;2016:455-464.
23. Lacson R, Harris K, Brawarsky P, et al. Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry. *J Digit Imaging*. 2015;28(5):567-575. doi:10.1007/s10278-014-9762-4
24. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Assessing the Utility of Automatic Cancer Registry Notifications Data Extraction from Free-Text Pathology Reports. *AMIA Annu Symp Proc*. 2015;2015:953-962.
25. Wieneke AE, Bowles EJA, Cronkite D, et al. Validation of natural language processing to extract breast cancer pathology procedures and results. *J Pathol Inform*. 2015;6. doi:10.4103/2153-3539.159215
26. Martinez D, Pitson G, MacKinlay A, Cavedon L. Cross-hospital portability of information extraction of cancer staging information. *Artificial Intelligence in Medicine*. 2014;62(1):11-21. doi:10.1016/j.artmed.2014.06.002
27. Coden A, Savova G, Sominsky I, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*. 2009;42(5):937-949. doi:10.1016/j.jbi.2008.12.005
28. Leroy G, Gu Y, Pettygrove S, Galindo MK, Arora A, Kurzius-Spencer M. Automated Extraction of Diagnostic Criteria From Electronic Health Records for Autism Spectrum Disorders: Development, Evaluation, and Application. *Journal of Medical Internet Research*. 2018;20(11):e10497. doi:10.2196/10497
29. Jackson RG, Patel R, Jayatilleke N, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*. 2017;7(1):e012012. doi:10.1136/bmjopen-2016-012012
30. Topaz M, Lai K, Dowding D, et al. Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application. *International Journal of Nursing Studies*. 2016;64:25-31. doi:10.1016/j.ijnurstu.2016.09.013
31. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306(8):848-855. doi:10.1001/jama.2011.1204
32. Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events. *J Am Med Inform Assoc*. 2011;18(4):491-497. doi:10.1136/amiainjnl-2011-000187
33. Pruitt P, Naidech A, Van Ornam J, Borczuk P, Thompson W. A natural language processing algorithm to extract characteristics of subdural hematoma from head CT reports. *Emerg Radiol*. January 2019. doi:10.1007/s10140-019-01673-4
34. Zech J, Pain M, Titano J, et al. Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports. *Radiology*. 2018;287(2):570-580. doi:10.1148/radiol.2018171093
35. Patterson OV, Freiberg MS, Skanderson M, J. Fodeh S, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovasc Disord*. 2017;17. doi:10.1186/s12872-017-0580-8

36. 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. Seattle, Washington, USA: ACM Press; 2012:1277. doi:10.1145/2145204.2145396





# 7

## Natural language processing for automated quantification of brain metastases reported in free-text radiology reports

---

Joeky T. Senders, Aditya V. Karhade, David J. Cote, Alireza Mehrtash,  
Nayan Lamba, Aislyn Dirisio, Ivo S. Muskens, William B. Gormley,  
Timothy R. Smith, Marike L.D. Broekman, Omar Arnaout

JCO CLIN CANCER INFORM. 2019 APR;3:1-9

# **Abstract**

## **Introduction**

Although the bulk of patient-generated health data is increasing exponentially, its utilization is impeded because most data comes in unstructured format, namely free-text clinical reports. A variety of natural language processing (NLP) methods have emerged to automate the processing of free text ranging from statistical to deep learning-based models; however, the optimal approach for medical text analysis remains to be determined. The aim of this study was to provide a head-to-head comparison of novel NLP techniques and inform future studies about their utility for automated medical text analysis.

## **Methods**

Magnetic resonance imaging reports of patients with brain metastases treated in two tertiary centers were retrieved and manually annotated using a binary classification (single metastasis versus two or more metastases). Multiple bag-of-words and sequence-based NLP models were developed and compared after randomly splitting the annotated reports into a training and test set in an 80:20 ratio.

## **Results**

A total of 1479 radiology reports of patients diagnosed with brain metastases were retrieved. The LASSO regression model demonstrated the best overall performance on the hold-out test set with an area under the receiver operating curve of 0.92 (95%CI 0.89–0.94), accuracy of 83% (95%CI 80–87%), calibration intercept of -0.06 (95%CI -0.14–0.01), and calibration slope of 1.06 (95%CI 0.95–1.17).

## **Conclusion**

Among various NLP techniques, the bag-of-words approach combined with a LASSO regression model demonstrated the best overall performance in extracting binary outcomes from free-text clinical reports. This study provides a framework for the development of machine learning-based NLP models, as well as a clinical vignette in patients diagnosed with brain metastasis.

# Introduction

In recent years, the volume and complexity of patient-generated health data are increasing exponentially. Although this data has the potential to propel clinical research in, its utilization is impeded because most of it comes in unstructured format, namely free-text clinical reports. Manual chart review remains therefore inevitable to identify patients and extract features of interest; however, as data sets are growing in size and granularity, this manual chart review becomes increasingly inefficient and even prone to error.

Natural language processing (NLP) is a branch of artificial intelligence that focuses on enabling computers to process human language. This technique could therefore facilitate clinical research in this patient population by accelerating the throughput of free-text clinical reports.<sup>1</sup> A variety of NLP approaches has emerged ranging from statistical to deep learning-based models; however, the optimal approach for automating the analysis of free-text medical documents remains to be determined.

The aim of this study was to provide a head-to-head comparison of NLP techniques for biomedical text analysis. Therefore, we have trained, evaluated, and compared various NLP techniques on their ability to process brain magnetic resonance imaging (MRI) reports of patients diagnosed with brain metastasis and quantify the number of metastases present. Although the current study focuses on radiology reports and brain metastasis patients, it provides a framework for development of NLP models for automated medical text analysis.

# Methods

## Participants

The Research Patient Data Registry (RPDR), which is a centralized clinical data registry maintained across the Partners Healthcare Hospitals Brigham and Women's Hospital and Massachusetts General Hospital, was queried for patients with known cerebral metastases using the international classification of diseases ninth revision (ICD-9) code 198.3. Patients were included if they had a radiological diagnosis of cerebral metastases and a complete free-text radiology report of the initial MRI brain examination. No follow-up reports were used as the number of lesions documented in these reports might have been distorted by treatment effects. This study was approved by the Institutional Review Board of Partners Healthcare, which waived the need for informed consent due to the retrospective nature of the study.

## Ground truth

The selected reports were randomized into six blocks. Each block was manually reviewed and annotated by two independent medical students (AK, DC, NL, AD) for the number of metastases present by means of a binary classification (single metastasis versus two or more metastases). Each student reviewer was blinded to the label generated by the other reviewer, and no additional clinical information apart from the text within the radiology report was provided. Conflicts in labeling were resolved by a final reviewer (JS, IM). Consensus in student classification was used to provide accurate labels for the training and test data, but also to replicate the way chart reviews are performed in clinical research. Although clinicians are commonly considered as most appropriate for collecting clinical data, recent studies suggest that the reliability of data collected by research assistants is not inferior, especially for information with low clinical complexity.<sup>2,3</sup>

## Development of a natural language processing model

The goal of this project was to compare various NLP approaches on their ability to classify MRI brain reports into those that describe a single metastasis versus those that describe multiple metastases. The approaches and algorithms used for this purpose can be classified into two broad categories: a bag-of-words and sequence-based approach. The bag-of-words approach considers the relative frequency of words in a document but ignores the order of these words. Similarly, the algorithms trained according to the bag-of-words approach in this project (logistic regression, least absolute shrinkage and selection operator [LASSO] regression, and multilayer perceptron) ignore the order of the words as well. Due to the rapid developments in the artificial neural network field, deep learning architectures have emerged that can model spatial or temporal configurations of the input features, which allow for a sequence-based NLP approach. These algorithms consider, for example, if words are close or far away from each other in the document. In this study, algorithms trained and evaluated according to a sequence-based approach included 1D-convolutional neural networks, Long-Short Term Memory (LSTM), and Gated Recurrent Unit (GRU).

## Preprocessing

The analysis of free-text reports required both generic and approach-specific preprocessing steps as described in Table 1. Free-text reports were cleaned from redundant or duplicate information (e.g., time, date, radiologist's signature, and white spaces between paragraphs), and stemming was used to teach the algorithm the equivalency between words with a similar lexical root and further reduce the vocabulary. These steps result in the most parsimonious representation of the lexical meaning in a text report.

Additional preprocessing steps for the bag-of-words approach included the n-gram technique and term frequency-inverse document frequency (TF-IDF) vectorization.<sup>4,5</sup> Because the bag-of-words approach ignores the order of the words, important word combinations can be missed. N-grams were, therefore, constructed to join adjacent word combinations and give them unique value and meaning. Distinct words such as 'midline' and 'shift' can, for example, be combined into the bigram 'midline\_shift'. The use of mono-, bi-, and trigrams was included as hyperparameter during cross-validation. The TF-IDF vectorization converts the text document into an array of numbers that reflects the frequency of words in the document relative to the frequency of these words across all documents.

An embedding layer was created for all sequence-based algorithms. In the embedding layer, a word can be represented by a vector of numbers instead of a single number. These numbers represent the coordinates of the word in the word embedding space. The words 'man', 'woman', 'boy', and 'girl' could, for example, be located in the same plane in the word embedding space but separated by dimensions related to gender and age. Word embedding therefore allows for the mapping of lexical relationships between individual words, and thus the statistical properties of a language.<sup>6</sup> The embedding layer was trained on the training set in a supervised fashion using a single perceptron as output node.

**TABLE 1.** Generic and algorithm specific preprocessing steps.

Preprocessing step	Explanation	Example
<b>Generic preprocessing</b>		
Raw text report	Unprocessed raw text reports	"...Exam is somewhat limited secondary to motion artifact. There is a 3.5 x 3.1 x 3.1 cm (TV by AP by CC) heterogeneously, predominantly peripherally enhancing mass centered within the right frontal lobe (series 13 image 87, series 14 image 9), which corresponds to the mass lesion identified on the recent CT 1/22/2010..."
Cleaning	Removal of redundant information (e.g., date, time, radiologist's signature, white spaces between sections, punctuation between letters, and stop words) and transformation to lower case letters.	"...exam somewhat limited secondary motion artifact 3.5 x 3.1 x 3.1 cm tv ap cc heterogeneously predominantly peripherally enhancing mass centered within right frontal lobe series 13 image 87 series 14 image 9 corresponds mass lesion identified recent ct..."
Stemming	Words with a similar lexical root are converged to the same stem word. For example, 'heterogeneously' and 'heterogeneity' are both converged to 'heterogen'.	"...exam somewhat limit secondari motion artifact 3.5 x 3.1 x 3.1 cm tv ap cc heterogen predominantli peripher enhanc mass center within right frontal lobe seri 13 imag 87 seri 14 imag 9 correspond mass lesion identifi recent ct..."
<b>Preprocessing for bag-of-words models*</b>		
N-gram construction	Adjacent individual word tokens were combined in mono-, bi-, and/or trigrams. In the example on the right, the stemmed report is converted to mono- and bi-grams.	"...exam exam_somewhat somewhat somewhat_limit limit_limit_limit_secondari secondari_secondari_motion motion motion_motion_artifact artifact_artifact_3.5 3.5 3.5_x x x_3.1 3.1 3.1_x x x_3.1 3.1 3.1_cm cm cm_tv tv_tv_ap ap_ap_cc cc_cc_heterogen heterogen..."
TF-IDF word vectorization	The relative frequency of individual word tokens in each document was calculated. Each document is represented by a vector, in which each number corresponds with the relative frequency of a certain grams in the document.	[0.08497, 0.06189, 0.06895, 0.06642, 0.05214, 0.05105, 0.08855, 0.11227, 0.15729, 0.06813, 0.06677, 0.05419, 0.05193, 0.06535, 0.06875, 0.07164, 0.13677, 0.08250, 0.06798, 0.09174, ...]
<b>Preprocessing for sequence-based models**</b>		
Embedding layer	An 8-dimensional embedding layer was trained and added as the first layer of each model. Each word in the document is represented by an 8-dimensional vector.	[[0.12, 0.28, 0.14, 0.48, 0.98, 0.77, 0.21, 0.87], [0.79, 0.66, 0.49, 0.49, 0.56, 0.39, 0.32, 0.51], [0.54, 0.33, 0.84, 0.72, 0.34, 0.47, 0.12, 0.42], ...]

Abbreviations: 1D=one dimensional; GRU=gated recurrent unit; LASSO=least absolute shrinkage and selection operator; LSTM=long-short term memory; TF-IDF=term frequency-inverse document frequency.

\*Logistic regression, LASSO-regression, and multi-layer perceptron

\*\*1D-convolutional neural networks, LSTM, and GRU

## Training and evaluation

The total data set was divided into a training and hold-out test set in an 80:20 ratio. Five-fold cross-validation was performed on the training set to optimize the hyperparameter settings. The final models were evaluated on the hold-out test set, which had not been used for preprocessing and hyperparameter tuning in any form. The output of the NLP models can be a predicted probability (between 0 and 1) or binary prediction (yes or no). Based on the type of output, the performance of the classification was captured in several parameters, including the area under the receiver operating curve (AUC), accuracy, and calibration.<sup>7</sup> The AUC is a measure of discrimination and represents the probability that an algorithm will rate cases higher than non-cases when two observations are chosen at random. Accuracy represents the percentage of reports classified correctly when the output of the model is binary. Logistic regression was considered as a benchmark for comparison with all other algorithms. The agreement between the predicted probabilities and the observed prevalence was visually assessed in a calibration plot and numerically assessed according to the calibration intercept and slope. A calibration intercept of 0 and slope of 1 is considered as perfect calibration. The NLP models were developed and evaluated in Python version 3.6 (Python Software Foundation, <http://www.python.org>) using the Keras and Scikit-learn libraries.<sup>8,9</sup> The difference in AUC was evaluated by means of the DeLong test and the difference in accuracy by means of the Chi-square test in R version 3.3.3 (R Core Team, Vienna, Austria, <https://cran.r-project.org>). The Benjamini-Hochberg procedure was used to correct for multiple testing. To promote the transparency and reproducibility of our work, we have deployed the source code on a publicly accessible GitHub repository ([https://github.com/jtsenders/nlp\\_brain\\_metastasis](https://github.com/jtsenders/nlp_brain_metastasis)). Additionally, a pseudocode is provided in Supplementary Table S1, which can be used to guide similar work in other clinical applications. The datasets generated and analyzed in the current study are available from the corresponding author on request.

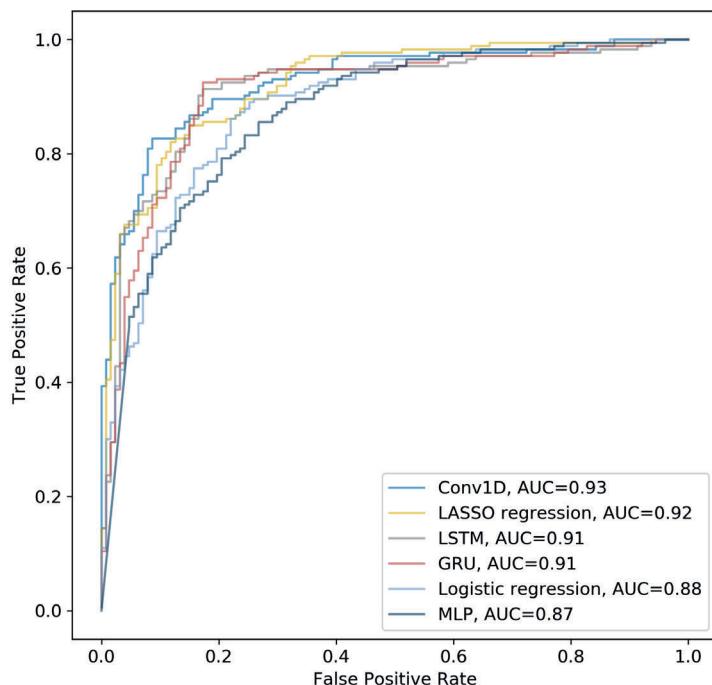
## Results

A total of 1479 reports of patients treated in one of the two Partners Hospitals was extracted by the RPDR query and eligible for inclusion in the current study. The annotated reports were divided into a training and hold-out test set of 1179 (79.7%) and 300 (20.3%) patients, respectively. The mean discordance rate between individual reviewers was 36.2%.

The AUCs on the hold-out test set of all six algorithms ranged between 0.87 and 0.93 (Figure 1), and the overall accuracies ranged between 64% and 87% (Table 2). By AUC,

the 1D-convolutional neural network demonstrated the best performance, which was significantly better compared to logistic regression (0.93 versus 0.88,  $p = 0.02$ ). LSTMs demonstrated the best performance in terms of overall accuracy, which was significantly better compared to logistic regression (87% versus 64%,  $p < 0.001$ ). The calibration across all models varied widely, and only the multilayer perceptron, GRU, and LASSO regression models included the intercept and slope values for perfect calibration in their confidence intervals (Figure 2) (Table 3).

Human annotation of the hold-out test set was completed in 6 days, whereas the best algorithm required 39.6 milliseconds for training, after which it could classify the entire hold-out test set in less than 0.8 milliseconds on a Central Processing Unit with four cores (2.2 GHz Intel Core i7).



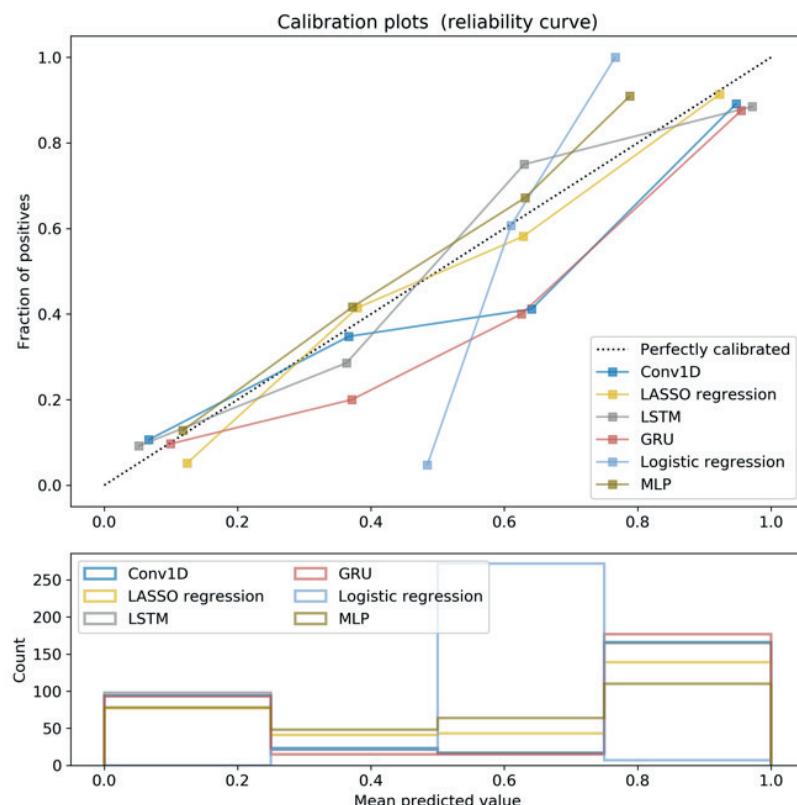
**FIGURE 1.** Receiver operating curves for all natural language processing models. Abbreviations: AUC=area under the receiver operating curve; Conv1D: one-dimensional convolutional neural network; GRU: gated recurrent unit; LSTM=long-short term memory; MLP=multilayer perceptron.

**TABLE 2.** Model performance according to the area under the receiver operating curve and accuracy, compared to logistic regression as benchmark.

Model	AUC (95%-CI)	p*	Accuracy (95%-CI)	p*
1D-convolutional neural network	0.93 (0.90–0.95)	0.02	85 (81–88)	<0.001
LASSO regression	0.92 (0.89–0.94)	0.02	83 (80–87)	<0.001
LSTM	0.91 (0.88–0.94)	0.12	87 (84–90)	<0.001
GRU	0.91 (0.88–0.93)	0.18	86 (82–89)	<0.001
Logistic regression	0.88 (0.85–0.92)	-	64 (60–68)	-
Multilayer perceptron	0.87 (0.84–0.90)	0.36	80 (76–83)	<0.001

Abbreviations: 1D=one dimensional; AUC=area under the receiver operating curve; CI=confidence interval; LASSO=least absolute shrinkage and selection operator; LSTM=long-short term memory; GRU=gated recurrent unit.

\*Corrected for multiple testing using the Benjamini-Hochberg procedure.



**FIGURE 2.** Calibration plot for all natural language processing models. Abbreviations: Conv1D: one-dimensional convolutional neural network; GRU: gated recurrent unit; LSTM=long-short term memory; MLP=multilayer perceptron.

**TABLE 3.** Model performance according to the calibration slope and intercept. A calibration intercept of 0 with a calibration slope of 1 is considered as perfect calibration.

<b>Model</b>	<b>Calibration</b>	
	<b>Slope (95%-CI)</b>	<b>Intercept (95%-CI)</b>
1D-convolutional neural network	0.90 (0.81 – 1.00)	0.03 (-0.04 – 0.09)
LASSO regression	1.06 (0.95 – 1.17)	-0.06 (-0.14 – 0.01)
LSTM	0.86 (0.78 – 0.95)	0.05 (-0.02 – 0.11)
GRU	0.92 (0.83 – 1.02)	-0.02 (-0.09 – 0.05)
Logistic regression	4.57 (3.94 – 5.20)	-2.19 (-2.57– -1.80)
Multilayer perceptron	1.14 (0.98 – 1.29)	-0.01 (-0.10 – 0.08)

Abbreviations: 1D=one dimensional; AUC=area under the receiver operating curve; CI=confidence interval; LASSO=least absolute shrinkage and selection operator; LSTM=long-short term memory; GRU=gated recurrent unit.

## Discussion

NLP constitutes a subfield of artificial intelligence that focuses on enabling computers to understand and process human languages.<sup>1</sup> Machine learning is another branch of artificial intelligence that focuses on enabling computer algorithms to learn from experience.<sup>10</sup> At their intersection, NLP harnessed with machine learning algorithms can learn how to process language by training on a vast number of labeled examples.<sup>11</sup> Among various NLP methods, the bag-of-words approach combined with a LASSO regression model demonstrated the best overall performance in extracting an equally-distributed, binary outcome from free-text clinical reports.

NLP has already been explored for the analysis of radiology reports of brain tumor patients, as well as other cancer types. Cheng et al. used NLP to analyze free-text radiology reports for tumor status classification.<sup>12</sup> Their NLP model had 80.6% sensitivity and 91.6% specificity in determining whether tumors had progressed, regressed, or remained stable. NLP for the analysis of radiology reports has also been explored in the context of other cancer types including hepatocellular carcinomas,<sup>13–16</sup> breast cancer,<sup>17–20</sup> lung cancer,<sup>21–23</sup> and other abdominal or pelvic tumors.<sup>11,24–27</sup> All studies that provided sufficient insight into their modeling approach utilized a bag-of-words approach. To our knowledge, the current study presents the first sequence-based NLP approach for the analysis of free-text radiology reports in oncology patients, as well as the first head-to-head comparison of sequence-based and bag-of-words models for medical text analysis.

## Limitations

Several limitations of the current study should be mentioned, which underline common barriers in NLP and machine learning modeling. Labels are necessary for the training and testing of algorithms, and human classification remains key for label generation in NLP tasks. Human classification, however, remains prone to error as well, which underlines the ambiguity of free-text clinical reports and the need for well-trained NLP models. In this study, a consensus in human classification was used as ground truth, which is a commonly used method to generate an approximation in the absence of actual ground truth.<sup>28</sup> Furthermore, this concept is already implemented in some frequently used machine learning algorithms, where the majority vote of many weak classifiers (e.g., decision tree) can result in a single strong classifier (e.g., random forest) referred to as ensemble learning.<sup>29</sup> In the current study, the complete data set was classified manually to generate labels for training and testing. However, when an NLP model will be put to use, only a minor portion will be labeled manually to predict the labels on the remaining data set. Due to the absence of labels in the remaining data set, external validation may not be feasible, and cross-validation remains the best approximation of model performance. Lastly, models trained on single institutional data might not generalize well to data from external institutions. Rather than deploying ready-to-use models, the current study therefore presents a framework for the development of NLP models that supports the overarching goal of automating the analysis of free-text clinical reports.

## Implications

Medical jargon can be heterogeneous in nature and expressed in various formats ranging from pathology and radiology reports to operative and discharge notes. This subset of unstructured data nonetheless follows a similar set of reporting norms, and thus statistical principles, which radically distinguishes this from human language used in newspapers, legal documents, or social media.<sup>30</sup> Although the current study focuses on brain metastasis patients and radiology reports, it can serve as a proof-of-concept for NLP of medical text. Therefore, the bag-of-words approach combined with a LASSO regression model may have a strong potential for NLP in other patient populations, clinical reports, and outcome measures as well. However, the nature of the NLP task of interest should align with the one used in the current study: extracting an equally-distributed, concrete binary outcome from free-text clinical reports. Within these boundaries, the presented NLP framework has the potential to facilitate retrospective clinical research by accelerating retrospective case identification and data extraction.

LASSO regression demonstrated superior overall performance among the bag-of-words models and 1D-convolutional neural networks among the sequence-based models. Although their preprocessing and analytical approaches differ, both algorithms provide strong methods for regularization to avoid overfitting.<sup>31, 32</sup> LASSO regression encourages simple models by penalizing the use of many coefficients, and convolutional layers extract higher-level features by applying filters on local regions of the input. Regularization is a key concept in machine learning and appears to be vital for both bag-of-words and sequence-based approaches in the current NLP task as well.<sup>29</sup> Although sequence-based approaches harnessed with recurrent and convolutional neural network architectures demonstrated higher overall performance than most bag-of-word approaches, their resultant models lacked the interpretability of regression-based algorithms, demanded longer training and prediction times, and required more careful hyperparameter tuning.

When constructing an NLP model, the choice of algorithm should be guided by the nature of the NLP task. If the NLP model should be fast, interpretable, and effective on a range of problems without tedious hyperparameter tuning, a bag-of-words approach based on a LASSO regression algorithm can be ideal.<sup>31</sup> If the order of the words is important, as with follow-up notes over time or higher-level relationships across distinct paragraphs, sequence-based approaches might be preferential.<sup>33,34</sup> Similarly, the metric of performance should align the overarching goal as close as possible. For example, sensitivity can be the metric of choice when comprehensiveness is the goal, and false-positives are more acceptable. On the other hand, specificity might be preferred when predicted cases should not be diluted with non-cases, and when false-negatives are more acceptable.

Future research should externally validate the current findings, thereby exploring and comparing the utility of bag-of-words and sequence-based NLP modeling in various patient populations, clinical reports, and outcome measures. In the current study, supervised learning methods were evaluated to investigate the utility of NLP for data extraction of unambiguous outcomes; however, future studies can also focus on extracting higher-lever concepts, such as the patient's survival probability or perception of quality of life. Although it remains questionable to what extent NLP can extract this information from clinical reports, it has the potential to pick up undetected patterns related to these outcomes. Furthermore, exploring the use of unsupervised learning in the absence of a prespecified outcome of interest might help in identifying natural, yet unknown clusters within the data. Lastly, future studies should consider the implications of automated medical text analysis parallel to the development of these techniques. NLP has the potential to increase the scale and velocity at which data

sets can be assembled, labelled, and analyzed; however, the increase in efficiency can come at the cost of transparency. Lack of transparency incurs the risk of large-scale misinterpretations of automatically assembled data sets. Researchers should therefore balance the yield of automated medical text analysis against the risk and consequences of potential misclassification. Establishing standards for model evaluation, as well as a minimal threshold for model performance, might help in estimating and mitigating this risk. Although the heterogeneity across NLP endeavors in healthcare might limit the establishment of uniform standards, defining general guidelines that can be further specified at study-level can foster a safe and effective implementation of NLP in medical research and even clinical care.

## **Conclusion**

The recent advent and widespread popularization of electronic medical records have led to an unprecedented volume of free-text clinical reports available for research purposes. Machine learning algorithms enable NLP techniques to learn from previously classified examples, thereby making it unnecessary to hard-code the rules for text analysis. Combining these techniques can therefore facilitate clinical research by optimizing the speed, accuracy, and consistency of clinical chart review. This study compares several NLP approaches for the classification of free-text radiology reports of brain metastasis patients, which can serve as a proof-of-concept and framework for NLP of electronic medical records.

## **Supplementary material**

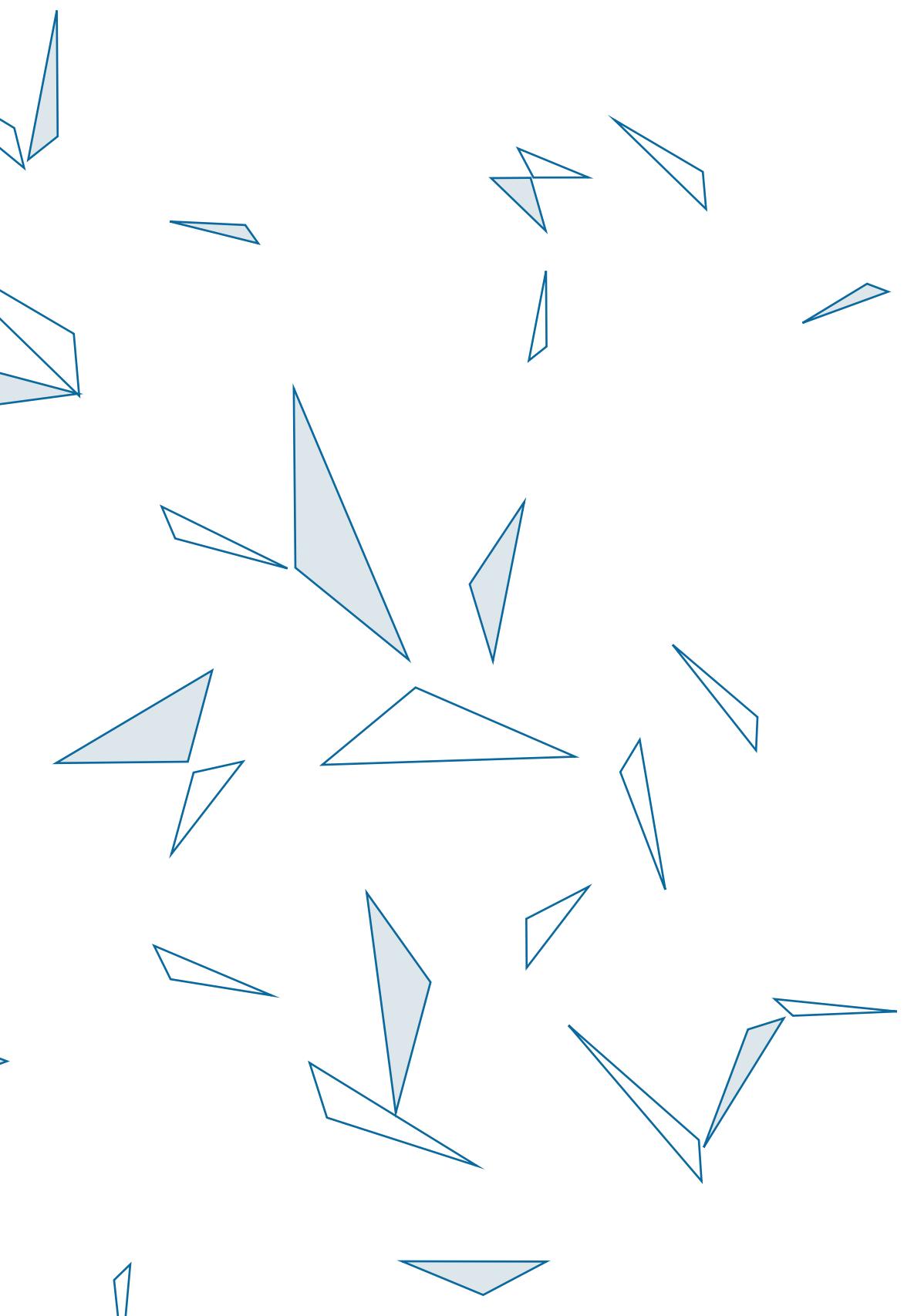
Supplementary Table S1 available online at:

[https://ascopubs.org/doi/10.1200/CCl.18.00138?url\\_ver=Z39.88-2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub%20%20pubmed](https://ascopubs.org/doi/10.1200/CCl.18.00138?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%20pubmed)

## References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc.* 2011;18(5):544-551. doi:10.1136/amiajnl-2011-000464
2. Mi MY, Collins JE, Lerner V, Losina E, Katz JN. Reliability of medical record abstraction by non-physicians for orthopedic research. *BMC Musculoskelet Disord.* 2013;14:181. doi:10.1186/1471-2474-14-181
3. Cruz CO, Meshberg EB, Shofer FS, McCusker CM, Chang AM, Hollander JE. Interrater Reliability and Accuracy of Clinicians and Trained Research Assistants Performing Prospective Data Collection in Emergency Department Patients With Potential Acute Coronary Syndrome. *Annals of Emergency Medicine.* 2009;54(1):1-7. doi:10.1016/j.annemergmed.2008.11.023
4. Nguyen VH, Nguyen HT, Duong HN, Snasel V. n-Gram-Based Text Compression. *Comput Intell Neurosci.* 2016;2016. doi:10.1155/2016/9483646
5. Jiang H, Li P, Hu X, Wang S. An improved method of term weighting for text classification. In: *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems.* Vol 1. ; 2009:294-298. doi:10.1109/ICICISYS.2009.5357842
6. Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent Word Embeddings of Free-Text Radiology Reports. *AMIA Annu Symp Proc.* 2018;2017:411-420.
7. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology.* 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
8. Deep Learning for Humans. Contribute to Keras-Team/Keras Development by Creating an Account on GitHub. Keras; 2018. <https://github.com/keras-team/keras>. Accessed October 28, 2018.
9. Scikit-Learn: Machine Learning in Python. Contribute to Scikit-Learn/Scikit-Learn Development by Creating an Account on GitHub. scikit-learn; 2018. <https://github.com/scikit-learn/scikit-learn>. Accessed October 28, 2018.
10. Obermeyer Z, Emanuel Ej. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med.* 2016;375(13):1216-1219. doi:10.1056/NEJMmp1606181
11. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *J Digit Imaging.* 2018;31(2):178-184. doi:10.1007/s10278-017-0027-x
12. Cheng LTE, Zheng J, Savova GK, Erickson BJ. Discerning Tumor Status from Unstructured MRI Reports—Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing. *J Digit Imaging.* 2010;23(2):119-132. doi:10.1007/s10278-009-9215-7
13. Sada Y, Hou J, Richardson P, El-Serag H, Davila J. Validation of Case Finding Algorithms for Hepatocellular Cancer from Administrative Data and Electronic Health Records using Natural Language Processing. *Med Care.* 2016;54(2):e9-e14. doi:10.1097/MLR.0b013e3182a30373
14. Yim W, Denman T, Kwan SW, Yetisgen M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Jt Summits Transl Sci Proc.* 2016;2016:455-464.
15. Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. *J Biomed Inform.* 2013;46(5):869-875. doi:10.1016/j.jbi.2013.06.014
16. Ping X-O, Tseng Y-J, Chung Y, et al. Information Extraction for Tracking Liver Cancer Patients' Statuses: From Mixture of Clinical Narrative Report Types. *Telemedicine and e-Health.* 2013;19(9):704-710. doi:10.1089/tmj.2012.0241
17. Bozkurt S, Gimenez F, Burnside ES, Gulsesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. *J Biomed Inform.* 2016;62:224-231. doi:10.1016/j.jbi.2016.07.001

18. Lacson R, Andriole KP, Prevedello LM, Khorasani R. Information from Searching Content with an Ontology Utilizing Toolkit (ISCOUP). *J Digit Imaging*. 2012;25(4):512-519. doi:10.1007/s10278-012-9463-9
19. Carrell DS, Halgrím S, Tran D-T, et al. Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence. *Am J Epidemiol*. 2014;179(6):749-758. doi:10.1093/aje/kwt441
20. Sippo DA, Warden GI, Andriole KP, et al. Automated Extraction of BI-RADS Final Assessment Categories from Radiology Reports with Natural Language Processing. *J Digit Imaging*. 2013;26(5):989-994. doi:10.1007/s10278-013-9616-5
21. Farjah F, Halgrím S, Buist DSM, et al. An Automated Method for Identifying Individuals with a Lung Nodule Can Be Feasibly Implemented Across Health Systems. *EGEMS (Wash DC)*. 2016;4(1). doi:10.13063/2327-9214.1254
22. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated Identification of Patients With Pulmonary Nodules in an Integrated Health System Using Administrative Health Plan Data, Radiology Reports, and Natural Language Processing. *Journal of Thoracic Oncology*. 2012;7(8):1257-1262. doi:10.1097/JTO.0b013e31825bd9f5
23. Wadia R, Akgun K, Brandt C, et al. Comparison of Natural Language Processing and Manual Coding for the Identification of Cross-Sectional Imaging Reports Suspicious for Lung Cancer. *JCO Clinical Cancer Informatics*. 2018;(2):1-7. doi:10.1200/CCI.17.00069
24. Pershad Y, Govindan S, Hara AK, et al. Using Naïve Bayesian Analysis to Determine Imaging Characteristics of KRAS Mutations in Metastatic Colon Cancer. *Diagnostics*. 2017;7(3):50. doi:10.3390/diagnostics7030050
25. Sevenster M, Bozeman J, Cowhy A, Trost W. Automatically Pairing Measured Findings across Narrative Abdomen CT Reports. *AMIA Annu Symp Proc*. 2013;2013:1262-1271.
26. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated Extraction of Grade, Stage, and Quality Information From Transurethral Resection of Bladder Tumor Pathology Reports Using Natural Language Processing. *JCO Clinical Cancer Informatics*. 2018;(2):1-8. doi:10.1200/CCI.17.00128
27. Gregg JR, Lang M, Wang LL, et al. Automating the Determination of Prostate Cancer Risk Strata From Electronic Medical Records. *JCO Clinical Cancer Informatics*. 2017;(1):1-8. doi:10.1200/CCI.16.00045
28. Valizadegan H, Nguyen Q, Hauskrecht M. Learning Classification Models from Multiple Experts. *J Biomed Inform*. 2013;46(6):1125-1135. doi:10.1016/j.jbi.2013.08.007
29. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
30. Wulff HR. The language of medicine. *J R Soc Med*. 2004;97(4):187-188.
31. Ranstam J, Cook JA. LASSO regression. *BJS*. 2018;105(10):1348-1348. doi:10.1002/bjs.10895
32. Rios A, Kavuluru R. Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles. *ACM BCB*. 2015;2015:258-267. doi:10.1145/2808719.2808746
33. Gehrman S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13(2). doi:10.1371/journal.pone.0192360
34. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evid Based Ment Health*. 2017;20(3):83-87. doi:10.1136/eb-2017-102688



# 8

## Deep learning for natural language processing of free- text pathology reports A comparison of learning curves

---

Joeky T. Senders, David J. Cote, Alireza Mehrtash, Robert Wiemann, William B. Gormley, Timothy R. Smith, Marike L.D. Broekman, Omar Arnaout

BMJ INNOVATIONS, 2020 VOL 6, ISSUE 4

# **Abstract**

## **Introduction**

Although clinically derived information could improve patient care, its full potential remains unrealized because most of it is stored in a format unsuitable for traditional methods of analysis, free-text clinical reports. Various studies have already demonstrated the utility of natural language processing algorithms for medical text analysis. Yet, evidence on their learning efficiency is still lacking. This study aimed to compare the learning curves of various algorithms and develop an open-source framework for text mining in healthcare.

## **Methods**

Deep learning and regressions-based models were developed to determine the histopathological diagnosis of brain tumor patients based on free-text pathology reports. For each model, we characterized the learning curve and the minimal required training examples to reach the area under the curve (AUC) performance thresholds of 0.95 and 0.98.

## **Results**

In total, we retrieved 7000 reports of 5242 brain tumor patients (2316 with glioma, 1412 with meningioma, and 1514 with cerebral metastasis). Conventional regression and deep learning-based models required 200-400 and 800-1500 training examples to reach the AUC performance thresholds of 0.95 and 0.98, respectively. The deep learning architecture developed in the current study required 100 and 200 examples, respectively, corresponding to a learning capacity that is two to eight times more efficient.

## **Conclusions**

This open-source framework enables the development of high-performing and fast learning natural language processing models. The steep learning curve can be valuable for contexts with limited training examples (e.g., rare diseases and events or institutions with lower patient volumes). The resultant models could accelerate retrospective chart review, assemble clinical registries, and facilitate a rapid learning health care system.

# Introduction

Clinically-derived patient information is generally increasing in volume and granularity with the expansion and improvement of online medical record systems.<sup>1</sup> Although analysis of this information allows for the generation of knowledge to improve future patient care, its full potential remains unrealized because most of it is stored in an unstructured, free-text format unsuitable for traditional methods of analysis. Manual review is necessary to extract and structure relevant patient information. As data sets continue to grow, however, manual chart review becomes increasingly inefficient, inconsistent, and prone to error.<sup>2</sup>

Various studies have already demonstrated the utility of automated methods for the processing and analysis of free-text clinical reports.<sup>3-14</sup> These algorithms have the potential to assist in structuring the immense stream of free-text clinical information produced on a day-to-day basis. However, they are generally evaluated on a static text corpus of clinical reports with a fixed number of training examples, thereby lacking evidence on their learning capacity (i.e., the required number of examples to train high-performing models).<sup>3-14</sup> Yet, learning efficiency is instrumental, as the availability of training examples can be limited, and their labelling can be time consuming or expensive.

In the current study, we aimed to compare the learning curves of various natural language processing approaches and develop an open-source framework for clinical text mining. Therefore, we have developed models that determine the histological diagnosis of brain tumor patients based on free-text pathology reports. Furthermore, we used various training samples to compare the efficiency of each algorithm's learning curve.

# Methods

## Participants

This study was conducted and reported according to the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis Or Diagnosis (TRIPOD) statement.<sup>15</sup> The Institutional Review Board of Brigham and Women's Hospital approved the current study and waived the need for informed consent due to its retrospective, observational design. We included all patients who underwent an operation at our institution for a histopathologically confirmed diagnosis of glioma, meningioma, or brain metastasis between January 2002 and July 2018. Patients were

identified through a departmental database that registers all neurosurgical patients who undergo an operation within our department. To retrieve the associated free-text pathology reports through which the diagnosis was made, we cross-linked the patient identification number and date of surgery with the pathology reports in our centralized institutional clinical data registry. These free-text pathology reports were used as input data for the natural language processing model. Manual annotations on the histopathological diagnosis were provided by a clinical reviewer with over 20 years of experience (R.W.). These annotations were used as labels for the target outcome in a binary fashion (diagnosis of interest versus other diagnoses). Some patients underwent multiple operations, thereby providing multiple pathology reports and associated diagnosis labels in the analysis. The total text corpus was split at the patient-level into a training, validation, and hold-out test set according to a 2:1:1 ratio. The test set was kept separate until the final performance evaluation. Differences in baseline characteristics between the training, validation, and hold-out test set were compared by means of the Chi-square test, analysis of variance (ANOVA), and Kruskal-Wallis test depending on the nature and distribution of the baseline characteristics.

## Preprocessing

The algorithms used for this purpose can be classified into two broad categories, regression-based and neural network-based algorithms.<sup>16,17</sup> The regression-based algorithms utilized a bag-of-words/n-grams approach, thereby considering the relative frequency of words or adjacent word combinations in a document but ignoring their order.<sup>17</sup> Deep learning-based algorithms, on the other hand, modeled the order of the words and the semantic relationships among them, as well.<sup>16,18</sup>

The analysis of free-text pathology reports required both generic and approach-specific preprocessing steps. Pseudocode in a generalizable format is provided in Table 1. These preprocessing steps are required to compress the lexical content of free-text pathology reports to the most parsimonious representation and convert these reports to a numeric format that could be processed by a classification algorithm. Redundant or duplicate text (time, date, pathologist's signature, unnecessary white spaces etc.) was removed, and stemming was used to converge words with a similar lexical root.<sup>19</sup> For the regression-based algorithms, we used n-grams to assign unique value and meaning to adjacent word combinations.<sup>20</sup> The term frequency-inverse document frequency (TF-IDF) vectorization was used to convert each document into an array of numbers reflecting the relative frequency of these word or word combinations.<sup>21</sup> For the deep learning models, we tokenized and zero padded the documents to convert them to a numeric format with the same length.<sup>22</sup>

**TABLE 1.** Pseudocode of the current study in a generalizable format.

Step 1: Data importation and general preprocessing	A. Import data frame with three columns containing the patient identifier, group label, and original clinical report. B. In the original report column, subsequently a. remove all redundant information (date, time, physician's signature, white spaces between sections, punctuation between letters, and stop words) and transform all letters to lower case letters. b. remove all English stop words except 'no' and 'not' c. apply Porter stemming algorithm d. apply preprocessing steps C. Divide the stemmed reports at the patient-level into a training, validation, and test set in a 2:1:1 ratio.
Step 2: Hyperparameter optimization	A. Hyperparameter optimization by means of weighted bootstrapping on the training and validation set. Hyperparameter settings are further explained in Supplementary Table S1.
Step 3: Evaluate model performance on the hold-out test set.	A. Train final models with optimal hyperparameter settings on the training set with 100 bootstraps for each model and each training fraction. B. Compute the predicted probabilities in the residual hold-out test set. C. Calculate the pooled mean AUC with standard deviation for each model and training fraction. D. Plot the performance in AUC against the training sample size to visualize the resultant learning curve for each model.

Abbreviations: AUC=area under the receiver operating curve

## Development of a natural language processing model

To compare the learning curve of the distinct approaches, model performance was evaluated for each diagnosis with varying samples ranging between 25 and 3000 reports of the training set. A bootstrapping procedure was utilized to optimize the hyperparameter settings. We used bootstrapping with replacement to draw random samples from the parent training set and trained 'naïve' algorithms with every iteration.<sup>23</sup> As such, all resultant models were solely trained on the randomly drawn sample while ignoring the rest of the parent training set. This bootstrapping procedure provided an estimate of performance for each hyperparameter setting by pooling the performance estimates of the distinct, sample-based models. To account for the higher variability in performance in the smaller samples and preserve the computational reproducibility of this study, the number of bootstraps was inversely weighted according to the sample size and comprised the integer division between the size of the total training set ( $n=3000$ ) and the size of the training sample. For example, training with a sample of 25 reports was bootstrapped 120 times.

Logistic regression, least absolute shrinkage and selection operator (LASSO) regression, and deep learning models were developed and compared as classifiers.<sup>24</sup> For the regression-based models, the use of mono-, bi-, and trigrams, the size of the vocabulary

in the TF-IDF vectorizer, and L1 regularization were presented as hyperparameter settings. The following hyperparameters were optimized for the deep learning models: dimensionality of the embedding layer, dropout, kernel size, L1 regularization, L2 regularization, learning rate, max pooling window, number of convolutional layers, number of dense layers, number of filters in the convolutional layers, number of nodes in the dense layers, report length, type of optimizer, and vocabulary size of the tokenizer. Embedding and convolutional layers constitute the most instrumental layers in deep learning models used for natural language processing. In the embedding layer, a word can be represented by a vector of numbers instead of a single number.<sup>16</sup> These numbers represent the coordinates of a word in the embedding space and as such reflect the semantic relationships between individual words. Convolutional layers capture local interactions among nearby words by applying transformations with smaller one-dimensional filters on local regions of the input data.<sup>25</sup> Convolutional neural network (CNN) models are characterized by these layers and currently widely investigated because of their strong potential for image and text processing. Among the deep learning-based models, we therefore specifically compared the best performing CNN with non-convolutional neural network architectures. Explanations of the other hyperparameters are provided in Supplementary Table S1.

## Evaluating final model performance

Training of the final models and evaluation on the residual hold-out test set was bootstrapped 100 times for each training fraction and model. The predicted outcomes of the natural language processing models constituted a probability of belonging to a histopathological class. Therefore, model performance was measured according to the area under the receiver operating characteristic curve (AUC).<sup>26</sup> The AUC is a measure of discrimination and represents the probability that an algorithm will rate a randomly selected case (i.e., category of interest) higher than a randomly selected non-case (i.e., all other cases). Model performance was pooled and weighted across the histopathological subclasses and plotted against the size of the training sample to construct each algorithm's learning curve. Based on these learning curves, we determined the minimal size of the training sample required to reach the AUC performance thresholds of >0.95 and >0.98. All models were trained and evaluated in Python version 3.6 (Python Software Foundation, <http://www.python.org>) using the Scikit-learn libraries. Figures for the incremental model performance were made in R version 3.3.3 (R Core Team, Vienna, Austria, <https://cran.r-project.org/>).<sup>27</sup>

# Results

A total of 7000 pathology reports from 5242 patients were retrieved. Among these patients, 2316 (44.2%) were diagnosed with glioma, 1412 (26.9%) with meningioma, and 1514 with cerebral metastasis (28.9%). Baseline characteristics for the training, validation, and test set are provided in Table 2. A statistically significant difference ( $p=0.038$ ) was found in the mean age across the cohorts. This difference was deemed to be of little clinical significance (57.0 years in the training set versus 56.4 years in the validation set) and likely the result of the large cohort sizes.

**TABLE 2.** Cohort characteristics

Patient Characteristics	Description	Training set (n = 2621)		Validation set (n = 1310)		Test set (n = 1311)			<i>p</i>
		No.	%	No.	%	No.	%		
Patient age	<50	728	27.8	423	32.3	394	30.1	0.060	
	>70	519	19.8	236	18.0	251	19.1		
	50-70	1374	52.4	651	49.7	666	50.8		
	Mean ± SD	57.0 ± 14.4		56.4 ± 14.2		56.7 ± 14.7			0.038
Sex	Female	1474	56.2	720	55.0	738	56.3	0.716	
	Male	1147	43.8	590	45.0	573	43.7		
Reports per patient	Median [IQR]	1 [1 - 1]		1 [1 - 1]		1 [1 - 1]		0.683	
Histopathological diagnosis	Glioma	1142	43.6	574	43.8	600	45.8	0.697	
	Meningioma	712	27.2	362	27.6	338	25.8		
	Metastasis	767	29.3	374	28.5	373	28.5		

Abbreviations: IQR=interquartile range; No.=sample size; *p*=*p*-value; SD=standard deviation

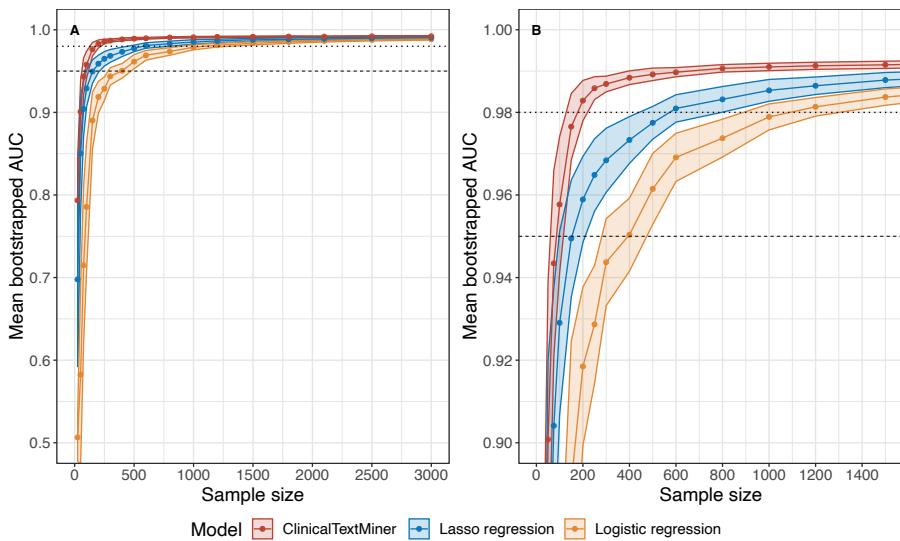
The neural network architecture developed in the current study, ClinicalTextMiner, demonstrated a steeper learning curve than regression-based models (Figure 1, Table 3) and other competing deep learning models (Figure 2). Regression-based algorithms required 200-400 and 800-1500 training examples to reach the AUC performance thresholds of 0.95 and 0.98, respectively. ClinicalTextMiner reached these thresholds with 100 and 200 examples, respectively, corresponding to a learning capacity that is two to eight times more efficient. The best performing CNN architecture reached the AUC performance threshold of 0.95 after training with at least 400 training examples and did not reach the performance threshold of 0.98. Furthermore, its

performance was less consistent compared to the other models as denoted by the larger standard deviations. Lastly, ClinicalTextMiner was the only model that reached the AUC performance threshold of >0.99, with 600 training examples.

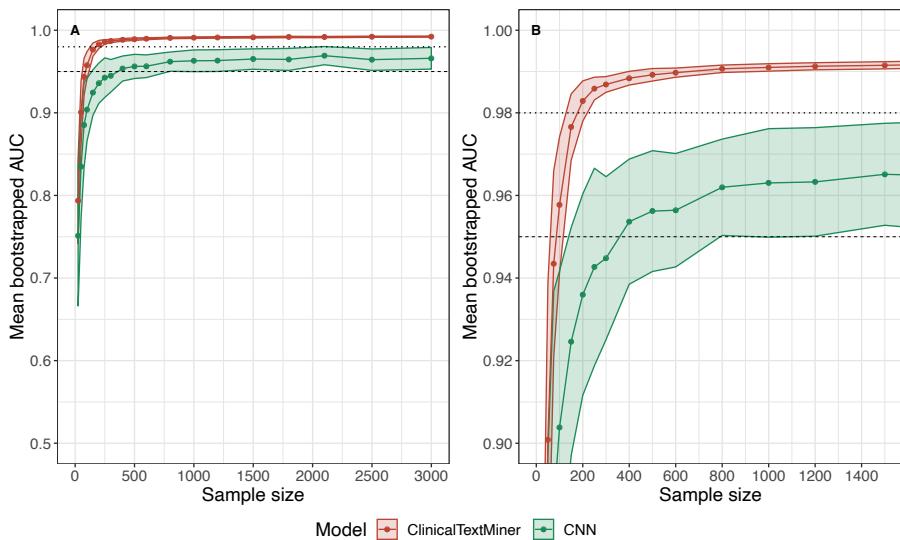
**TABLE 3.** Incremental model performance according to the area under the receiver operating characteristic curve.

Sample size	Bootstrapped AUC			
	Deep learning-based models		Regression-based models	
	ClinicalTextMiner	CNN	Lasso regression	Logistic regression
25	0.794±0.052	0.751±0.085	0.698±0.106	0.507±0.029
50	0.901±0.038	0.835±0.062	0.850±0.070	0.583±0.109
75	0.943±0.023	0.885±0.052	0.904±0.033	0.715±0.096
100	0.958±0.016	0.904±0.038	0.929±0.022	0.785±0.078
150	0.977±0.008	0.925±0.028	0.949±0.014	0.890±0.034
200	0.983±0.005	0.936±0.024	0.959±0.010	0.918±0.019
250	0.986±0.003	0.943±0.024	0.965±0.009	0.929±0.014
300	0.987±0.002	0.945±0.020	0.968±0.008	0.944±0.011
400	0.988±0.002	0.954±0.015	0.973±0.006	0.950±0.009
500	0.989±0.002	0.956±0.015	0.977±0.004	0.961±0.009
600	0.990±0.001	0.956±0.014	0.981±0.003	0.969±0.006
800	0.991±0.001	0.962±0.012	0.983±0.003	0.974±0.005
1000	0.991±0.001	0.963±0.013	0.985±0.003	0.979±0.003
1200	0.991±0.001	0.963±0.013	0.986±0.002	0.981±0.002
1500	0.992±0.001	0.965±0.012	0.988±0.002	0.984±0.002
1800	0.992±0.001	0.965±0.013	0.989±0.002	0.985±0.002
2100	0.992±0.001	0.969±0.011	0.989±0.002	0.986±0.001
2500	0.992±0.001	0.964±0.013	0.990±0.001	0.988±0.001
3000	0.992±0.001	0.966±0.013	0.990±0.001	0.989±0.001

Abbreviations: AUC=area under the receiver operating characteristic curve; CNN=convolutional neural networks



**FIGURE 1.** Incremental model performance comparing ClinicalTextMiner to regression-based algorithms according to the area under the receiver operating characteristic curve (A). Enlarged panel can be found on the right (B). The dashed line represents the 0.95 performance threshold and the dotted line represents the 0.98 performance threshold. Abbreviations: AUC= area under the receiver operating characteristic curve.



**FIGURE 2.** Incremental model performance comparing ClinicalTextMiner to the best performing convolutional neural network (CNN) model according to the area under the receiver operating characteristic curve (A). Enlarged panel can be found on the right (B). The dashed line represents the 0.95 performance threshold and the dotted line represents the 0.98 performance threshold. Abbreviations: AUC= area under the receiver operating characteristic curve; CNN=convolutional neural network.

## Discussion

In this study, we compared the learning curves of various natural language processing techniques for clinical text mining. We identified a deep learning architecture that learns two to eight times more efficiently than competing deep learning and regressions-based approaches. The underlying source code has been released on a publicly accessible repository, thereby allowing for external application, validation, and optimization.

Few other groups have explored the use of deep learning for natural language processing of free-text clinical reports.<sup>3-14</sup> Although these studies demonstrate the strong potential of deep learning for clinical text mining, this has only been demonstrated on a static text corpus of clinical reports with a fixed number of training examples. The current body of literature therefore still lacks evidence on the learning capacity of natural language processing as it applies to clinical text mining. To the best of our knowledge, this is the first study that compares the learning curve of deep learning and other competing algorithms in their ability to process free-text clinical reports. Additionally, by removing the convolutional layer and combining the embedding layer with strong methods of regularization, we identified a model architecture that learns more efficiently compared to competing algorithms and even CNNs, thereby requiring less training examples to develop high performing clinical text mining models.

## Limitations

Several limitations of the current study should be mentioned, which underline common barriers in natural language processing and machine learning modeling. The current models are trained on single institutional data and might not generalize well to data from external institutions, which may have different styles and routines in the language used in clinical reports. Instead of the resultant models, the underlying code pipeline was therefore made publicly accessible in order to promote the reproducibility and external generalizability of the current work. Labels were necessary for training and testing of the natural language processing models, and these were derived through manual chart review; however, manual chart review remains prone to error as well. In the current study, a trained clinical reviewer (R.W.) with over 20 years of experience has provided the required labels. The current study utilized a three-way classification problem (i.e., glioma, metastasis, or meningioma) to evaluate the learning curves. However, the number of diagnostic classes can vary widely dependent on the scope of interest, which could also impact the efficiency of the resultant learning curves.

## Implications

Despite these limitations, we believe the current study provides valuable insight into the learning capacity of various methods for clinical text mining, as well as an open-source framework for developing these tools. In the current project, we used natural language processing to extract the histopathological diagnosis from free-text pathology reports of brain tumor patients. This application was chosen because the histopathological diagnosis constitutes the cornerstone for patient grouping in clinical research and day-to-day patient care. Currently, retrospective case identification is often based on ICD-9 codes and manual chart review. The accuracy of ICD-9 codes is questionable because they are developed for billing purposes and are often registered by non-medically trained assistants. Several studies have investigated the utility of ICD-9 codes, and all reported poor performance in terms of accuracy, sensitivity, specificity, or positive predictive value, depending on the diagnosis of interest.<sup>28-31</sup> Manual chart review, on the other hand, is labor intensive and drastically limits the speed, scale, and consistency of retrospective case identification.

It remains unclear why ClinicalTextMiner was able to outperform traditional (bag-of-words/n-gram) and novel methods (convolutional neural networks) of natural language processing. An explanation could be that the bag-of-words/n-gram approach focuses primarily on the relative frequency of certain word or adjacent word combinations in the text, whereas ClinicalTextMiner models the semantic properties and relations as well, due to the embedding layer at its base. The strong methods of regularization (i.e., pooling and dropout) could potentially avoid overfitting to statistical noise, which is introduced when analyzing semantic properties of words in addition to the relative word frequencies.

The model developed in the current study required only 100 (i.e., 30-35 per diagnostic group) and 200 training examples (i.e., 60-70 per diagnostic group) to reach the AUC performance thresholds of 0.95 and 0.98, respectively. This learning capacity can be instrumental as it allows for the development of high-performing clinical text mining models in applications with limited training examples. As such, models can also be developed in hospitals with lower patient volumes or utilized in the context of rare diseases and events. Furthermore, it reduces the workload on clinical experts who have to label each training example manually. In addition to the histopathological diagnosis, this open-source framework can guide the development of models to extract various clinical concepts from other report types, simply by providing the appropriate reports

and training labels. For example, labelling a radiology report with the size and location of the lesion allows for the development of a model that can extract radiological characteristics.<sup>32–35</sup>

Natural language processing could improve clinical research and patient care in several ways. The automatic nature could accelerate retrospective chart review to an unprecedented scale and allow for the assembly of large 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. Lastly, by extracting patient characteristics and outcomes automatically, natural language processing could facilitate a health care system that continuously learns from clinically derived data. As such, these algorithms might assist in structuring the immense stream of free-text clinical information produced on a day-to-day basis. Models could, for example, be developed to automate trivial, administrative processes with low clinical impact (i.e., assigning the appropriate billing codes) or construct flagging systems and safety nets for matters of high-clinical impact (i.e., serious findings reported in diagnostic studies).

The current framework allows for the prediction of a single outcome and can be transformed to predict other outcomes as well. Developing distinct models for different outcomes in the same text corpus can, however, be computationally inefficient. After all, the underlying patterns and features learned from the text corpus might be generalizable to multiple outcomes and circumvent the need for a duplicate, time consuming training process. A solution could be to freeze the base model and repeat the training process only for the dense layers at the top of the network.<sup>36</sup> Another solution is emerging in the computational field and encompasses the development of deep learning algorithms for multiclass, multilabel classification problems. Future studies should therefore focus on evaluating the utility of these multiclass, multilabel algorithms for clinical text mining.<sup>37</sup> Furthermore, the learning curve of natural language processing algorithms remains relatively unexplored in the current literature and requires further investigation as well. The open-source framework developed in the current study could guide the construction of similar models for other clinical applications. Particularly, the efficient model architecture identified in the current study could be valuable for optimizing the hyperparameter settings in other deep learning-based natural language processing models. Lastly, future studies should not merely focus on the analytical challenges, but also on the implications of relying on automated methods for medical text analysis.

# Conclusion

We developed an open-source natural language processing pipeline that can determine the histopathological diagnosis of brain tumor patients based on free-text pathology reports. A deep learning architecture was identified that learns two to eight times more efficiently than competing deep learning and regressions-based approaches. This framework could facilitate clinical research by providing an automatic and deterministic method for medical text analysis.

## Supplementary material

Supplementary Table S1 and Figure S1 are available online at:

<https://innovations.bmjjournals.com/content/6/4/192>

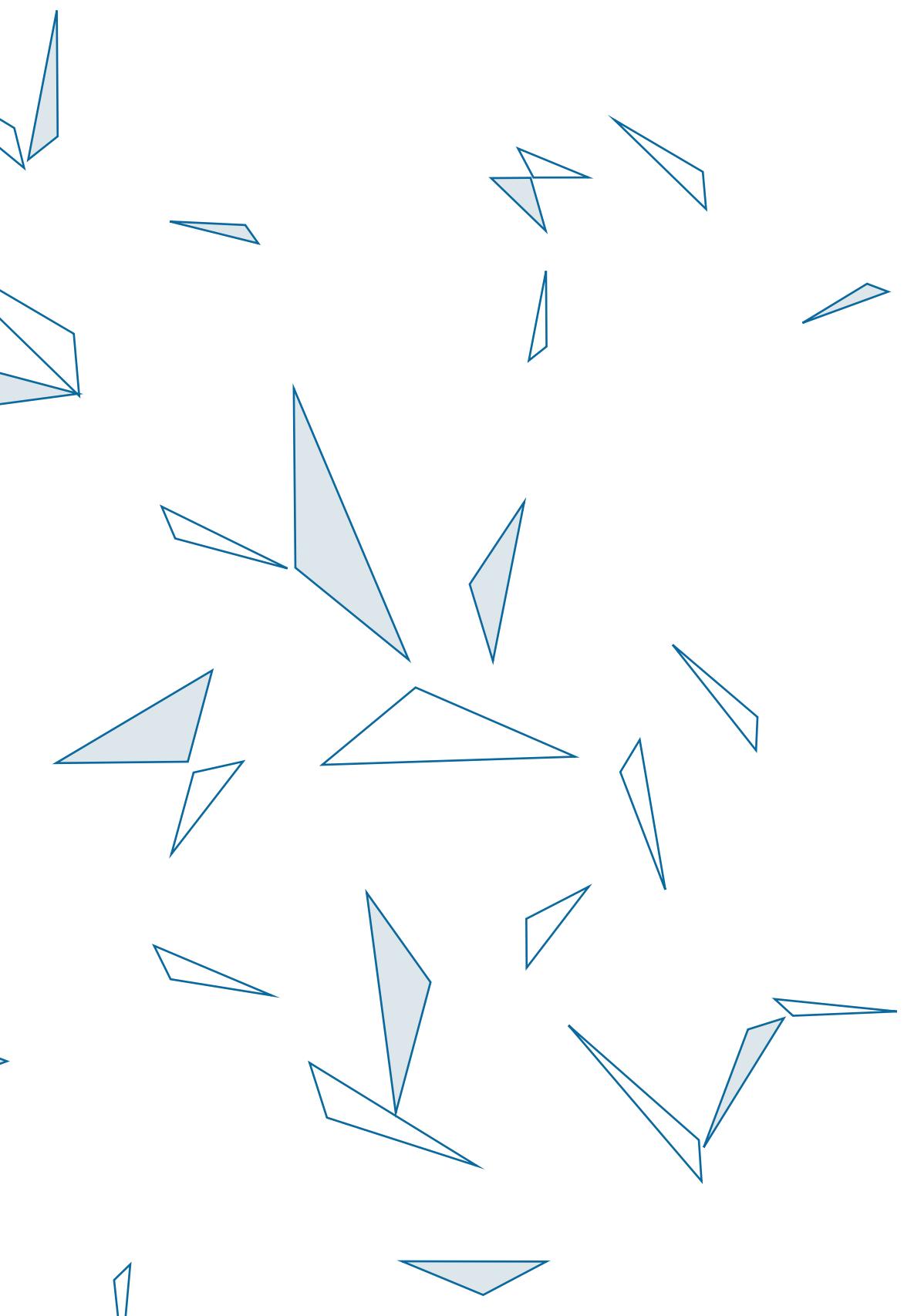
## References

1. Evans RS. Electronic Health Records: Then, Now, and in the Future. *Yearb Med Inform.* 2016;(Suppl 1):S48-S61. doi:10.15265/IYS-2016-s006
2. Matt V, Matthew H. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof.* 2013;10. doi:10.3352/jeehp.2013.10.12
3. Bao Y, Deng Z, Wang Y, et al. Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes. *JCO Clinical Cancer Informatics.* 2019;(3):1-9. doi:10.1200/CCl.19.00042
4. Senders JT, Karhade AV, Cote DJ, et al. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. *JCO Clinical Cancer Informatics.* 2019;In Press
5. Shi X, Yi Y, Xiong Y, et al. Extracting entities with attributes in clinical text via joint deep learning. *J Am Med Inform Assoc.* doi:10.1093/jamia/ocz158
6. Spandorfer A, Branch C, Sharma P, et al. Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. *Eur Radiol Exp.* 2019;3(1):37. doi:10.1186/s41747-019-0118-1
7. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *J Digit Imaging.* 2018;31(2):178-184. doi:10.1007/s10278-017-0027-x
8. Bacchi Stephen, Oakden-Rayner Luke, Zerner Toby, Kleining Timothy, Patel Sandy, Jannes Jim. Deep Learning Natural Language Processing Successfully Predicts the Cerebrovascular Cause of Transient Ischemic Attack-Like Presentations. *Stroke.* 2019;50(3):758-760. doi:10.1161/STROKEAHA.118.024124
9. Leyh-Bannurah S-R, Tian Z, Karakiewicz PI, et al. Deep Learning for Natural Language Processing in Urology: State-of-the-Art Automated Extraction of Detailed Pathologic Prostate Cancer Data From Narratively Written Electronic Health Records. *JCO Clinical Cancer Informatics.* 2018;(2):1-9. doi:10.1200/CCl.18.00080
10. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. *JAMA Netw Open.* 2018;1(6). doi:10.1001/jamanetworkopen.2018.3451
11. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology.* 2019;291(1):196-202. doi:10.1148/radiol.2018180921
12. Kehl KL, Elmarakeby H, Nishino M, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol.* Published online July 25, 2019. doi:10.1001/jamaoncol.2019.1800
13. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J Am Med Inform Assoc.* Published online May 28, 2019. doi:10.1093/jamia/ocz063
14. He T, Puppala M, Ezeana CF, et al. A Deep Learning-Based Decision Support Tool for Precision Risk Assessment of Breast Cancer. *JCO Clin Cancer Inform.* 2019;3:1-12. doi:10.1200/CCl.18.00121
15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594. doi:10.1136/bmj.g7594
16. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc.* 2020;27(3):457-470. doi:10.1093/jamia/ocz200
17. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev.* 2019;8. doi:10.1186/s13643-019-1074-9

18. Gonçalves S, Cortez P, Moro S. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*. Published online 2019. doi:10.1007/s00521-019-04334-2
19. Cai T, Giannopoulos AA, Yu S, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics*. 2016;36(1):176-191. doi:10.1148/rg.2016150080
20. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551. doi:10.1136/amiajnl-2011-000464
21. Zhang W, Yoshida T, Tang X. TFIDF, LSI and multi-word in information retrieval and text categorization. In: *2008 IEEE International Conference on Systems, Man and Cybernetics*; 2008:108-113. doi:10.1109/ICSMC.2008.4811259
22. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9(4):611-629. doi:10.1007/s13244-018-0639-9
23. Henderson AR. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clin Chim Acta*. 2005;359(1-2):1-26. doi:10.1016/j.cccn.2005.04.002
24. Ranstam J, Cook JA. LASSO regression. *BJS*. 2018;105(10):1348-1348. doi:10.1002/bjs.10895
25. Zola P, Cortez P, Ragno C, Brentari E. Social Media Cross-Source and Cross-Domain Sentiment Classification. *International Journal of Information Technology & Decision Making (IJITDM)*. 2019;18(05):1469-1499.
26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
27. Modern Optimization with R | Paulo Cortez | Springer. Accessed April 6, 2020. <https://www.springer.com/gp/book/9783319082622>
28. Labovitz DL. Accuracy and yield of ICD-9 codes for identifying children with ischemic stroke. Published online November 22, 2018. Accessed November 22, 2018. <http://n.neurology.org/content/accuracy-and-yield-icd-9-codes-identifying-children-ischemic-stroke>
29. Pimentel MA, Browne EN, Janardhana PM, et al. Assessment of the Accuracy of Using ICD-9 Codes to Identify Uveitis, Herpes Zoster Ophthalmicus, Scleritis, and Episcleritis. *JAMA Ophthalmol*. 2016;134(9):1001-1006. doi:10.1001/jamaophthalmol.2016.2166
30. Guevara RE, Butler JC, Marston BJ, Plouffe JF, File TM, Breiman RF. Accuracy of ICD-9-CM Codes in Detecting Community-acquired Pneumococcal Pneumonia for Incidence and Vaccine Efficacy Studies. *Am J Epidemiol*. 1999;149(3):282-289. doi:10.1093/oxfordjournals.aje.a009804
31. Goldstein Larry B. Accuracy of ICD-9-CM Coding for the Identification of Patients With Acute Ischemic Stroke. *Stroke*. 1998;29(8):1602-1604. doi:10.1161/01.STR.29.8.1602
32. Tang R, Ouyang L, Li C, et al. Machine learning to parse breast pathology reports in Chinese. *Breast Cancer Res Treat*. 2018;169(2):243-250. doi:10.1007/s10549-018-4668-3
33. Imler TD, Morea J, Kahi C, et al. Multi-Center Colonoscopy Quality Measurement Utilizing Natural Language Processing. *The American Journal of Gastroenterology*. 2015;110(4):543-552. doi:10.1038/ajg.2015.51
34. Imler TD, Morea J, Kahi C, Imperiale TF. Natural Language Processing Accurately Categorizes Findings From Colonoscopy and Pathology Reports. *Clinical Gastroenterology and Hepatology*. 2013;11(6):689-694. doi:10.1016/j.cgh.2012.11.035
35. Jouhet V, Defossez G, Burgun A, et al. Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer. *Methods Inf Med*. 2012;51(03):242-251. doi:10.3414/ME11-01-0005
36. Shin H-C, Roth HR, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*. 2016;35(5):1285-1298. doi:10.1109/TMI.2016.2528162

37. Gargiulo F, Silvestri S, Ciampi M. Deep Convolution Neural Network for Extreme Multi-label Text Classification:  
In: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*.  
SCITEPRESS - Science and Technology Publications; 2018:641-650. doi:10.5220/0006730506410650





# 9

## Summary



# SUMMARY

Despite improved surgical and adjuvant treatment options, survival in patients with a malignant brain tumor remains dismal. Over the past decades, the volume and complexity of clinically-derived patient data (i.e., imaging, genomics, free-text etc.) is increasing exponentially. Machine learning provides a vast range of algorithms that can learn from this data and guide clinical decision-making by providing accurate patient-level predictions. The current thesis describes several studies along the continuum of the machine learning spectrum as it applies to neurosurgical oncology.

## Part I: Outcomes and risk factors in neurosurgical oncology

This part characterizes postoperative outcomes and associated risk factors in patients undergoing craniotomy for a malignant brain tumor. In **Chapter 2**, we have analyzed a cohort of 7376 patients identified through the National Surgical Quality Improvement Program (NSQIP) registry. Among patients undergoing craniotomy for a primary malignant brain tumor, 12.9% experienced a major complication within 30 days after surgery, most of which occurred during the initial hospital stay. The most common postoperative major complications were reoperation (5.1%), venous thromboembolism (VTE, 3.5%), and death (2.6%). The most common reasons for reoperation and unplanned readmission were intracranial hemorrhage (18.5% of all re-operated patients) and wound-related complications (11.9% of all re-admitted patients). The American Society of Anesthesiologists (ASA)-classification and preoperative functional status were most frequently identified as predictors, as well as the strongest predictors of postoperative morbidity and mortality.

Due to the high incidence and substantial impact on postoperative morbidity, a subsequent in-depth analysis (**Chapter 3**) was performed in the same cohort to characterize the rates, timing, and predictors of VTE and intracranial hemorrhage. This study demonstrated that the increased risk of VTE extends beyond the period of hospitalization, especially for pulmonary embolism, whereas intracranial hemorrhages occurred predominantly within the first days after surgery. Although age and body mass index (BMI) were identified as overall predictors of VTE, distinct risk profiles were observed in patients who developed a pulmonary embolism versus deep venous thrombosis, as well as patients who developed a VTE in-hospital versus post-discharge.

Given the persistent risk of pulmonary embolism beyond hospitalization, as well as the accumulation of intracranial hemorrhage events in the first days after surgery, a single-institutional retrospective cohort study (**Chapter 4**) was performed to

investigate the influence of continuing prophylactic anticoagulation beyond discharge. In a multivariable analysis of 301 patients who underwent craniotomy for a high-grade glioma, the VTE rate in patients who received thromboprophylaxis for 21 days after surgery was not statistically significantly lower compared to patients who received thromboprophylaxis up to discharge. However, a significantly higher rate of intracranial hemorrhage was seen in the prolonged thromboprophylaxis group. Functional status and BMI were confirmed as predictors of VTE in this cohort.

## Part II: Predictive analytics in neurosurgical oncology

This part constitutes a study that evaluates the utility of machine learning for outcome prediction in neuro-oncology. **Chapter 5** compared 15 commonly used statistical and machine learning algorithms in their ability to predict survival in glioblastoma patients based on demographic, socio-economic, clinical, and radiological features. In a cohort of 20,281 patients, identified through the Surveillance Epidemiology and End Results (SEER) registry, the accelerated failure time model outperformed competing statistical and machine learning algorithms in terms of prediction performance (C-index = 0.70), as well as interpretability, predictive utility, and computational efficiency. This model was therefore deployed as a free, publicly-available online calculator.

## Part III: Natural language processing in neurosurgical oncology

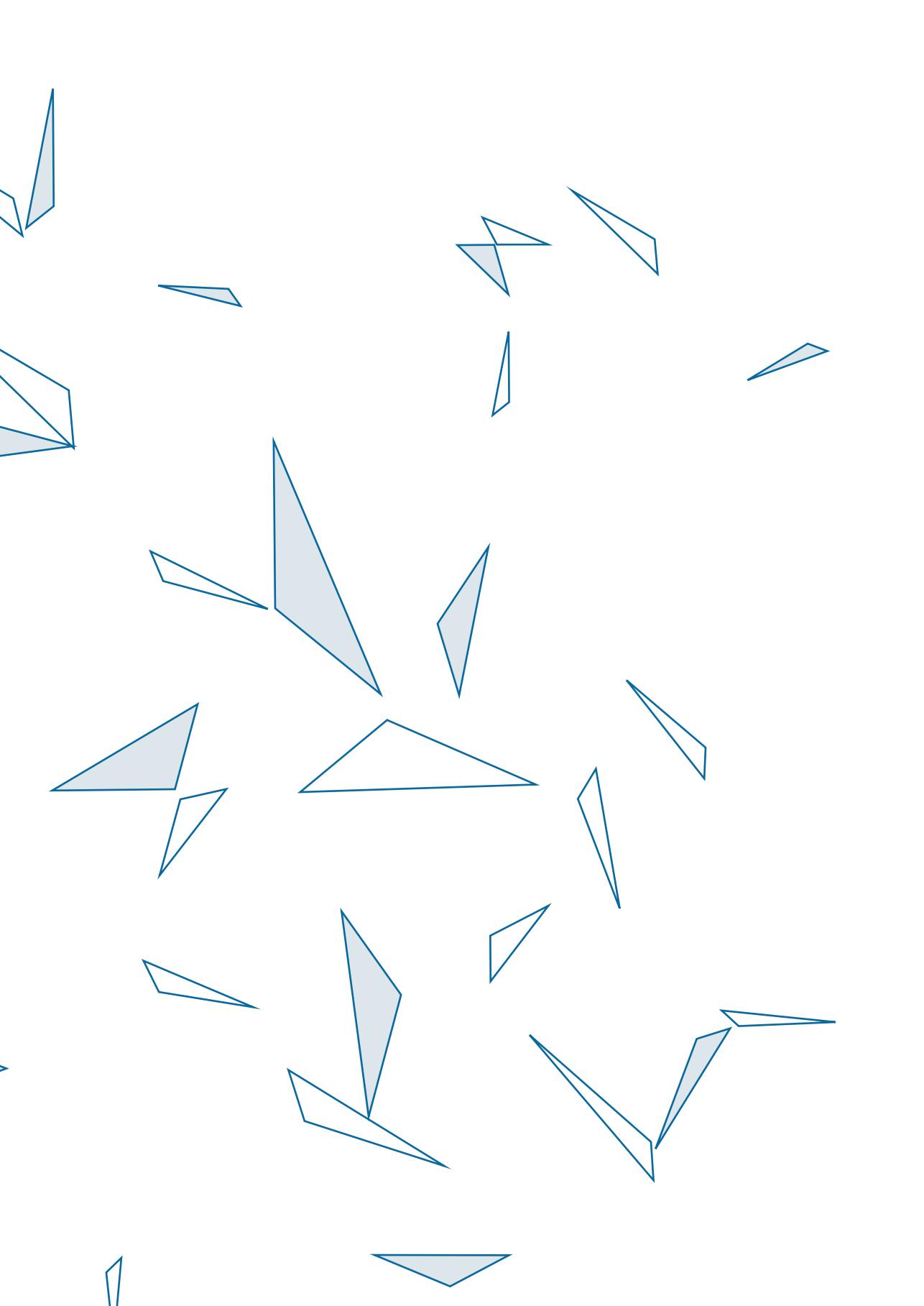
In this part, various natural language processing approaches were developed and evaluated for medical text analysis in brain tumor patients. In **Chapter 6**, we developed an open-source natural language processing framework for automating the extraction of clinical information. To this end, we analyzed a text corpus of narratively written radiology reports of glioblastoma patients ( $n = 562$  reports) with a regression-based algorithm (LASSO regression) as classifier. The resultant pipeline was able to extract 15 radiographic features with high to excellent performance (AUC 0.82-0.98). Model performance was correlated with the interrater agreement of the manually provided labels ( $p = 0.904$ ;  $p < .001$ ), but not with the frequency distribution of the variables of interest ( $p = 0.179$ ;  $p = .52$ ).

In the subsequent **Chapter 7**, we developed and 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 free-text radiology reports ( $n = 1,472$ ) of brain metastases patients into reports that describe

solitary versus multiple metastases. The bag-of-words approach combined with a LASSO regression algorithm still demonstrated to outperform competing algorithms in terms of model discrimination (AUC = 0.92) and calibration.

In **Chapter 8**, we examined 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 deep learning model (ClinicalTextMiner) equipped with stronger methods of regularization. In addition to modeling only the relative frequency of words (like LASSO regression does), the resultant model was able to model the semantic complexity of text documents as well, without overfitting to statistical noise. The number of required training samples to reach the predetermined performance thresholds (AUC of 0.95 and 0.98) was two to eight times lower for ClinicalTextMiner compared to regression and conventional deep learning-based architectures.

All custom computer code and software developed throughout this thesis have been made publicly-available to facilitate the transparency and reproducibility of this work. The associated URL-links can be found in the **Open-source code and software appendix**.



# 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.<sup>1</sup> 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,

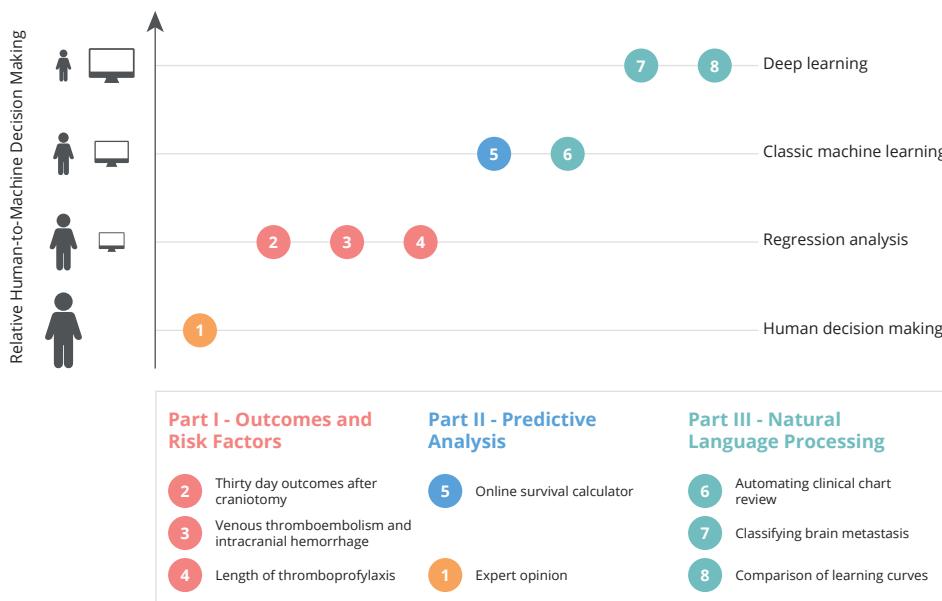
interactions to include, or transformation to perform if the variables are composed of individual words in a text document, pixels in a picture, genes on a chromosome, or voxels in an MRI scan, let alone specify all these model properties by hand?

## **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.<sup>2</sup> 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.<sup>3</sup> 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.<sup>4</sup> The current thesis describes several studies along the continuum of the machine learning spectrum as it applies to neurosurgical oncology (Figure 1).



**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 heterogeneous 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 reiterate 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.<sup>5,6</sup> 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.<sup>7</sup>

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.<sup>3</sup> 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.<sup>7</sup> 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.<sup>8</sup> 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.

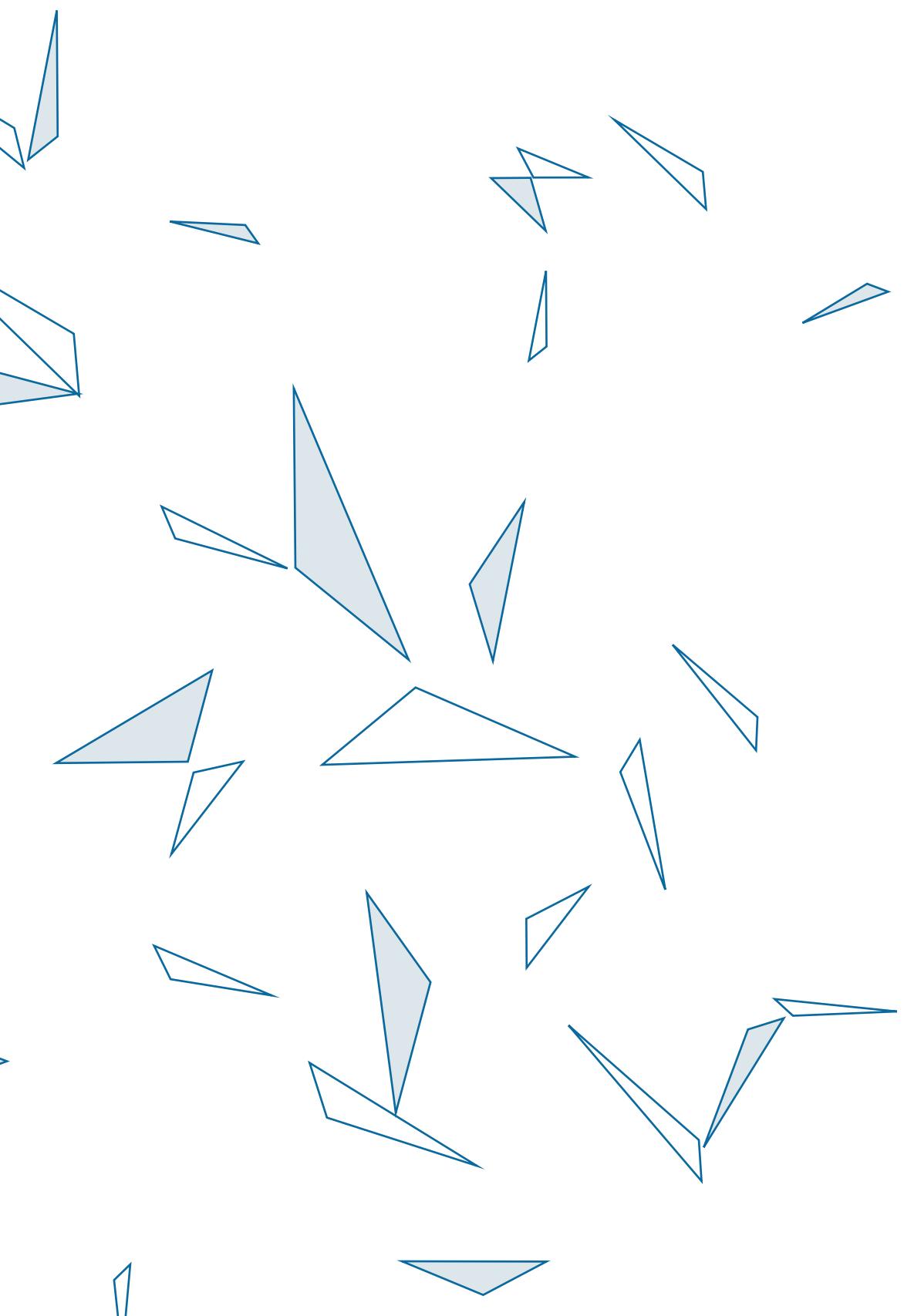
Lastly, machine learning provides powerful methods for mapping numeric input to numeric output. However, not everything is reducible to numbers, especially not in healthcare. Primitive clinical characteristics, pictures, text, and images can all be expressed as 0’s or 1’s and thus easily be incorporated 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>



# A

## Appendices

---

- Nederlandse samenvatting
- Open-source code and software
- Authors and affiliations
- Acknowledgements
- Report of scholarship
- About the author



# SAMENVATTING

Ondanks verbeterde chirurgische en adjuvante behandelopties blijft de overleving van patiënten met een maligne hersentumor somber. In de recente decennia is het volume en de complexiteit van medische informatie (i.e., radiologie, DNA, vrije tekst) exponentieel gegroeid. Machine learning biedt een breed scala aan algoritmen die kunnen leren van deze data om toekomstige patiëntenzorg te informeren en optimaliseren voor de individuele patiënt. Dit proefschrift onderzoekt verscheidene algoritmen verspreid over het continuüm van het machine learning spectrum en hun toepassing binnen de neurochirurgische oncologie.

## Deel I: Uitkomsten en risico factoren in de neurochirurgische oncologie

Dit deel onderzoekt de postoperatieve uitkomsten en geassocieerde risicofactoren in patiënten die een craniotomie ondergaan voor een maligne hersentumor. In **Hoofdstuk 2** hebben we een cohort van 7376 patiënten van de National Surgical Quality Improvement Program (NSQIP) database geanalyseerd. Van alle patiënten die een craniotomie ondergingen voor een primaire maligne hersentumor, ontwikkelde 12,9% een grote complicatie binnen 30 dagen na de operatie, waarvan het merendeel plaatsvond tijdens de initiële opname. De meest voorkomende postoperatieve grote complicaties waren reoperatie (5,1%), veneuze trombo-embolie (VTE, 3,5%) en mortaliteit (2,6%). De meest voorkomende redenen voor reoperatie en ongeplande heropname waren respectievelijk intracraniale bloedingen (18,5% van alle reoperaties) en wond-gerelateerde complicaties (11,9% van alle heropnames). De American Society of Anesthesiologists (ASA)-classificatie en de preoperatieve functionele status werden geïdentificeerd als de meest frequente en sterkste voorspellers van postoperatieve morbiditeit en mortaliteit.

Vanwege de hoge incidentie en substantiële impact op morbiditeit werd een vervolg analyse (**Hoofdstuk 3**) gedaan in hetzelfde cohort om de incidentie, timing, en risicofactoren van VTE en intracraniale bloedingen te onderzoeken. Deze studie liet zien dat het verhoogde risico voor VTE, en longembolieën in het bijzonder, persisteert to na de periode van opname. Daarentegen vonden de intracraniale bloedingen voornamelijk plaats in de eerste dagen na operatie. Hoewel leeftijd en body mass index (BMI) geïdentificeerd werden als algemene risicofactoren voor VTE, werden er verschillende risicotyperijlen gevonden voor een longembolie versus diep veneuze trombose en voor een trombo-embolisch event tijdens de opname versus na ontslag.

Vanwege het persisterende risico op een longembolie na de opname en de accumulatie van intracraniële bloedingen in de eerste dagen na chirurgie, werd een retrospectieve studie (**Hoofdstuk 4**) gedaan om de veiligheid en effectiviteit van het continueren van thromboprophylaxe na ontslag te onderzoeken. Een multivariate analyse van 301 patiënten liet zien dat thromboprophylaxe tot 21 dagen na de operatie niet geassocieerd was met een lagere VTE incidentie in vergelijking met thromboprophylaxe tot aan ontslag. Echter werden er meer intracraniële bloedingen gezien in de eerste groep. Functionele status en BMI werden in dit cohort bevestigd als voorspellers van VTE.

## Deel II: Predictieve analyse in de neurochirurgische oncologie

In dit deel, zijn verscheidene machine learning modellen ontwikkeld en geëvalueerd in hun vermogen om klinisch vergaarde patiënt informatie te analyseren en zodoende de klinische besluitvorming te informeren voor patiënten met een maligne hersentumor.

**Hoofdstuk 5** vergeleek 15 veelgebruikte statistische en machine learning algoritmen in hun vermogen om overleving te voorspellen in glioblastoom patiënten op basis van gestructureerde demografische, socio-economische, klinische en radiologische informatie. In een cohort van 20.281 patiënten van de Surveillance Epidemiology and End Results (SEER) database, was de accelerated failure time model in vergelijking met andere algoritmen superieur wat betreft voorspellend vermogen, maar ook wat betreft de interpreteerbaarheid, toepasbaarheid en computationale efficiëntie. Dit model is dan ook beschikbaar gesteld middels een gratis en publiek-toegankelijke, online software.

## Deel III: Natural language processing in de neurochirurgische oncologie

In dit deel, zijn verscheidene natural language processing methoden ontwikkeld voor de verwerking en analyse van medische tekstinformatie bij hersentumor patiënten. In **Hoofdstuk 6**, is een open-source natural language processing framework ontwikkeld om de extractie van medische informatie te automatiseren. Hiervoor hebben we een dataset van radiologieverslagen van glioblastoom patiënten ( $n = 562$ ) geanalyseerd middels een regressie (LASSO-regressie) algoritme. Het uiteindelijke framework was in staat om 15 radiologische karakteristieken te extraheren met een hoog tot excellent discriminerend vermogen (AUC 0,82 - 0,98). Dit discriminerend vermogen was gecorreleerd met de consensus tussen klinische experts ten aanzien van de manueel geconstrueerde labels ( $p = 0,904$ ;  $p < 0,001$ ), maar niet met de frequentieverdeling van de desbetreffende variabelen ( $p = 0,179$ ;  $p = 0,52$ ).

In **Hoofdstuk 7**, zijn verscheidene statistische (logistische en LASSO-regressie), klassieke machine learning (artificial neural networks), en deep learning (convolutional neural networks, gated recurrent unit, en long short-term memory) modellen ontwikkeld en vergeleken in hun vermogen om radiologieverslagen ( $n = 1.472$ ) van patiënten met een hersenmetastasen te kunnen classificeren in verslagen met solitaire versus multifocale metastasen. De bag-of-words benadering in combinatie met een LASSO-regressie algoritme overtrof de andere modellen op het gebied van discriminerend vermogen ( $AUC = 0,92$ ) en kalibratie.

In **Hoofdstuk 8**, hebben we de leercurves van verschillende algoritmen onderzocht voor het bepalen van de histopathologische diagnose van hersentumor patiënten op basis van de pathologieverslagen. In deze studie, hebben we een deep learning model (ClinicalTextMiner) ontwikkeld en gemodificeerd met strengere methoden voor regularisatie. Naast de relatieve frequentie van individuele woorden of woord combinaties (zoals bij LASSO-regressie), was dit model ook in staat om de semantische complexiteit van tekstdocumenten te modelleren, zonder te overfitten op statistische ruis. Zodoende had ClinicalTextMiner 2 tot 8 keer minder training voorbeelden nodig om de vooraf bepaalde afkapwaarde in discriminerend vermogen ( $AUC$  van 0,95 en 0,98) te behalen.

Alle code en software geschreven in dit proefschrift is openbaar beschikbaar om de transparantie en reproduceerbaarheid van dit werk te bevorderen. De URL-links kunnen gevonden worden in de **Open-source code and software appendix**.



# OPEN SOURCE CODE AND SOFTWARE

To facilitate the reproducibility and transparency of the current work, we have released all developed code and software from **Part II and Part III** on publicly-accessible servers or repositories. This allows for external validation and optimization in other institutions, patients, and outcomes. We strongly advocate future research to adopt the concept of open source coding to facilitate the implementation and acceptance of high-performing models in clinical care.

## Chapter 5

Title: An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning

Product: Online survival calculator

Link: <https://cnoc-bwh.shinyapps.io/gbmsurvivalpredictor/>

## Chapter 6

Title: Automating clinical chart review — an open-source natural language processing pipeline developed on free-text radiology reports of glioblastoma patients

Product: Source code

Link: [https://github.com/jtsenders/nlp\\_glioblastoma](https://github.com/jtsenders/nlp_glioblastoma)

## Chapter 7

Title: Natural language processing for automated quantification of brain metastases reported in free-text radiology reports

Product: Source code

Link: [https://github.com/jtsenders/nlp\\_brain\\_metastasis](https://github.com/jtsenders/nlp_brain_metastasis)

## Chapter 8

Title: Deep learning for natural language processing of free-text pathology reports — a comparison of learning curves

Product: Source code

Link: [https://github.com/jtsenders/nlp\\_learning\\_curves](https://github.com/jtsenders/nlp_learning_curves)



## AUTHORS AND AFFILIATIONS

### **Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA**

#### *Department of Neurosurgery*

Almekkawi, Ahmad Kareem

Ashby, Joanna L.

Calvachi, Paola.

Cho, Logan D.

Cote, David J.

Dawood, Hassan Y.

DiRisio, Aislyn

Goldhaber, Nicole H.

Gormley, William B.

Karhade, Aditya V.

Lamba, Nayan

McNulty, John J.

Schulte, Isabelle S.

Smith, Timothy R.

Wiemann, Robert

#### *Department of Radiology*

Mehrtash, Alireza

### **University Medical Center Utrecht, Utrecht University, Utrecht, NL**

#### *Department of Neurosurgery*

Robe, Pierre A.

van Bentum, G.

van Essen, M.

#### *Department of Neurology*

Snijders, Tom J.

Seute, Tatjana

#### *Department of Medical Oncology*

de Vos, Filip Y.F.L.

**Leiden University Medical Center, Leiden University, Leiden,  
NL**

*Department of Neurosurgery*

Muskens, Ivo S.

*Department of Neurology*

Taphoorn, Martin J.B.

**T.H. Chan Harvard School of Public Health, Boston, MA, USA**

*Department of Biostatistics*

Staples, P.

**Dana-Farber Cancer Institute**

*Department of Neurology*

Reardon, David A.

**Haaglanden Medical Center, The Hague, NL**

*Department of Neurosurgery*

Broekman, Marike L.D.

# ACKNOWLEDGEMENTS

Mr. Dr. Broekman; dear Marike, none of this work would have been possible without your guidance. You are an inexhaustible source of bright ideas, always driven by the pursuit of curiosity. Your passion for research is highly contagious. Thanks for all the lessons you have taught me, inviting me to Boston, and mentoring me throughout my development as researcher and doctor.

Prof. dr. Peul; dear Wilco, thanks for the opportunity to conduct my research in the context of a PhD program under your supervision. Your work has significantly impacted clinical care. Your ability to connect with people based on mutual empathy and respect inspires others and me to pursue the same.

Dr. Gormley, Dr. Smith, Dr. Arnaout, and Dr. Zaidi; thanks for providing the academic environment, guidance, and trust to translate my curiosity into scientific projects. Under your supervision, I have collaborated with a wide range of people with various expertise, yet all driven by the same intrinsic motivation.

Dr. Staples and Dr. Mehrtash; dear Patrick and Alireza, thanks for introducing me to the world of data science and teaching me the R and python programming languages.

Dr. Mekary, Hassan Dawood, David Cote, Aditya Karhade, Mark Zaki and Maya Harary and all other CNOc members; it was an absolute privilege to join this team of young bright minds. I learnt a wide range of skills simply by collaborating with all the talented people working at CNOc. Your willingness and enthusiasm to contribute to new ideas and projects constitutes the core of this highly productive research group.

All my neurosurgical colleagues at University Medical Center Utrecht; thanks for providing the opportunity and creating the stimulating environment to train as a neurosurgical resident. Looking forward to further expand my neurosurgical exposure and experience in the years to come.

Ivo Muskens and Enrico Martin; after countless Nero coffees and countless hours working side to side in the research lab, you both substantially contributed to the current thesis, as well as the unforgettable Boston experience accompanied with it.

Rens Varkevisser, Yassine Ochen, Stijn van Roessel, Stijn de Jonge, Sjors Klompmaker, Sabine Egeler, Thomas Bodewes, Noortje Hagemeijer, Hannah Harp and Lauren Denholfe from the Appleton 88, as well as Sophia É, Claudia Bargon, Robin Tan, Valerie ter Wengel, Reinier de Vries, Luc Eijkelestatam, Merel Stor, Sophie Zwanenburg,

Hassanin Alkaduhimi, Luis Baptist, Paul Ogink, Olivier and Vincent Groot and all my other Bostonian friends; although the current thesis only reflects the academic work performed in Boston, you guys have been responsible for the unforgettable memories created here.

Bas Evers, Bastiaan van Hoorn, Daan Donkers, Floris-Jan Pieters, Geert Kronemeijer, Guido Houben, Kaz de Bruijn, Maarten Buitenhuis, Maurits Barendregt, Rafael van der Steen, Roan Vlutters, Sander Genissen, and Thierry van Benten; a diverse group of friends and personalities characterized by mutual appreciation and knowing how to put things in perspective. Thanks for providing the dynamic social environment in Utrecht and Amsterdam.

Bob Vonk, Bob Logjes, David Toledo, Frits en Jelle van Schooten, Salo Oof, en Stijn van Roessel; thanks for the inspirational discussion groups fueled with focus and creativity.

Stijn Franssen and Hans van Lennep; although we still pretend to grow up, I am grateful to see little has changed since we became friends as eleven-year-olds.

Bastiaan van Hoorn, Bob Logjes, and Thijs Kuiper; the greatest and most memorable experiences in the past decade all started as random ideas in one of our minds. You guys have always stimulated and joined me in experiencing life and the world to the fullest extent.

Pap, Mam and Opa; thanks for your unconditional love and support. You have raised me with a sense of positivity that enabled me to dream big and dare to fail. Although the path I have chosen is very different from yours, your support and understanding still helps me with each and every step on the way.

Tim Senders; dear brother, once the two of us were planning on going into the circus together as acrobats. Along the way, we have discovered and experienced a multitude of equally extraordinary adventures. Now, only one of us has made it into the circus and the other one is making successful television shows. I still look up to you as my older brother and best friend.

Puck Nijssen, dear Puck, you are a whirlwind of positive energy, as well as a loving and caring person with a natural ability to connect with others. Thanks for entering my life in the last chapters of this thesis. Looking forward to add many more chapters with you in my life.

# REPORT OF SCHOLARSHIP

## List of peer-reviewed publications

- Deep learning for natural language processing of free-text pathology reports — a comparison of learning curves.  
**Senders JT**, Cote DJ, Mehrtash A, Wiemann R, Gormley WB, Smith TR, Broekman MLD, Arnaout O. *BMJ Innov.* 2020 vol 6, issue 4.
- Orally Administered 5-aminolevulinic Acid for Isolation and Characterization of Circulating Tumor-Derived Extracellular Vesicles in Glioblastoma Patients.  
Maas SLN, van Solinge TS, Schnoor R, Yekula A, **Senders JT**, de Vrij J, Robe P, Carter BS, Balaj L, Arkesteijn GJA, Nolte-'t Hoen ENM, Broekman MLD. *Cancers (Basel).* 2020 Nov 7.
- Survival prediction of glioblastoma patients—are we there yet? A systematic review of prognostic modeling for glioblastoma and its clinical potential.  
Tewarie IA, **Senders JT**, Kremer S, Devi S, Gormley WB, Arnaout O, Smith TR, Broekman MLD. *Neurosurg Rev.* 2020 Nov 6.
- Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review.  
Groot OQ, Bongers MER, Ogink PT, **Senders JT**, Karhade AV, Bramer JAM, Verlaan JJ, Schwab JH. *Clin Orthop Relat Res.* 2020 Jul 30
- Beta-blockers and glioma: a systematic review of preclinical studies and clinical results.  
Tewarie IA, **Senders JT**, Hulsbergen AFC, Kremer S, Broekman MLD. *Neurosurg Rev.* 2020 Mar 14.
- Preoperative functional MRI use in neurooncology patients: a clinician survey.  
Stopa BM, **Senders JT**, Broekman MLD, Vangel M, Golby AJ. *Neurosurg Focus.* 2020 Feb 1;48(2):E11.
- Automating clinical chart review: an open-source natural language processing pipeline developed on free-text radiology reports from patients with glioblastoma.  
**Senders JT**, Cho LD, Calvachi P, McNulty JJ, Ashby JL, Schulte IS, Almekkawi AK, Mehrtash A, Gormley WB, Smith TR, Broekman MLD, Arnaout O. *JCO Clin Cancer Inform.* 2020 Jan;4:25-34.

- An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning.  
**Senders JT**, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, Gormley WB, Smith TR, Broekman ML, Arnaout O. *Neurosurgery*. 2020 Feb 1;86(2):E184-E192.
- Passive data collection and use in healthcare: a systematic review of ethical issues.  
Maher NA, **Senders JT**, Hulsbergen AFC, Lamba N, Parker M, Onnela JP, Bredenoord AL, Smith TR, Broekman MLD. *Int J Med Inform*. 2019 Sep;129:242-247.
- Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bi-dimensional measurement.  
Chang K, Beers AL, Bai HX, Brown JM, Ly KI, Li X, **Senders JT**, Kavouridis VK, Boaro A, Su C, Bi WL, Rapalino O, Liao W, Shen Q, Zhou H, Xiao B, Wang Y, Zhang PJ, Pinho MC, Wen PY, Batchelor TT, Boxerman JL, Arnaout O, Rosen BR, Gerstner ER, Yang L, Huang RY, Kalpathy-Cramer J. *Neuro Oncol*. 2019 Jun 13. pii: noz106.
- Trends in high-impact neurosurgical randomized controlled trials published in general medical journals: a systematic review.  
Karhade AV, **Senders JT**, Martin E, Muskens IS, Zaidi HA, Broekman ML, Smith TR. *World Neurosurg*. 2019 May 17. pii: S1878-8750(19)31357-9.
- Routine blood tests do not predict survival in glioblastoma patients – multivariable analysis of 497 patients.  
Maas SLN, Draisma, Snijders TJ, **Senders JT**, Berendsen S, Seute T, Schiffelers RM, Solinge van WW, ten Berg MJ, Robe PA, Broekman MLD. *World Neurosurg*. 2019 Mar 14. pii: S1878-8750(19)30688-6.
- Natural language processing for automated quantification of brain metastases reported in free-text radiology reports.  
**Senders JT**, Karhade AV, Cote DJ, Mehrtash A, Lamba N, DiRisio A, Muskens IS, Gormley WB, Smith TR, Broekman MLD, Arnaout O. *JCO Clin Cancer Inform*. 2019 Apr;3:1-9.
- Randomized controlled trials comparing surgery to non-operative management in neurosurgery: a systematic review.  
Martin E, Muskens IS, **Senders JT**, DiRisio AC, Karhade AV, Zaidi HA, Moojen WA, Peul WC, SmitTR, Broekman MLD. *Acta Neurochir (Wien)*. 2019 Feb 23.

- Machine learning reveals multimodal MRI patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low- and high-grade gliomas.  
Zhou H, Chang K, Bai HX, Xiao B, Su C, Bi WL, Zhang PJ, **Senders JT**, Vallières M, Kavouridis VK, Boaro A, Arnaout O, Yang L, Huang RY. *J Neurooncol.* 2019 Jan 19.
- International practice variation in postoperative imaging of chronic subdural hematoma patients.  
Hulsbergen AFC, Yan SC, Stopa BM, DiRisio A, **Senders JT**, van Essen MJ, van der Burgt SME, Smith TR, Gormley WB, Broekman MLD. *J Neurosurg.* 2018 Dec 21:1-8.
- Treatment and survival of osteosarcoma and Ewing sarcoma of the skull: a SEER-database analysis.  
Martin E, **Senders JT**, Ter Wengel PV, Smith TR, Broekman MLD. *Acta Neurochir (Wien).* 2019 Feb;161(2):317-325.
- Behavior and attitudes among European neurosurgeons - An international survey.  
Muskens IS, van der Burgt SME, **Senders JT**, Lamba N, Peerdeeman SM, Broekman ML. *J Clin Neurosci.* 2018 Sep;55:5-9.
- Information-based medicine in glioma patients: a clinical perspective.  
**Senders JT**, Harary M, Stopa BM, Staples P, Broekman MLD, Smith TR, Gormley WB, Arnaout O. *Comput Math Methods Med.* 2018 Jun 13.
- Timing of surgery in traumatic brachial plexus injury: a systematic review.  
Martin E, **Senders JT**, DiRisio AC, Smith TR, Broekman MLD. *J Neurosurg.* 2018 May 1:1-13.
- Length of thromboprophylaxis in patients operated on for a high-grade glioma: A retrospective study.  
**Senders JT**, Snijders TJ, van Essen M, van Bentum GM, Seute T, de Vos FY, Smith TR, Robe PA, Broekman MLD. *World Neurosurg.* 2018 Jul.
- A nationwide analysis of 30-day adverse events, unplanned readmission, and length of hospital stay after peripheral nerve surgery in extremities and the brachial plexus.  
Martin E, Muskens IS, **Senders JT**, Cote DJ, Smith TR, Broekman MLD. *Microsurgery.* 2019 Feb;39(2).

- Clinical challenges of glioma and pregnancy: a systematic review.  
van Westrhenen A, **Senders JT**, Martin E, DiRisio AC, Broekman MLD. *J Neurooncol.* 2018 Aug;139(1):.
- Thirty-day outcomes after craniotomy for primary malignant brain tumors: a national surgical quality improvement program analysis.  
**Senders JT**, Muskens IS, Cote DJ, Goldhaber NH, Dawood HY, Gormley WB, Broekman MLD, Smith TR. *Neurosurgery*. 2018 Dec 1;83(6).
- Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging.  
Chang K, Bai HX, Zhou H, Su C, Bi WL, Agbodza E, Kavouridis VK, **Senders JT**, Boaro A, Beers A, Zhang B, Capellini A, Liao W, Shen Q, Li X, Xiao B, Cryan J, Ramkissoon S, Ramkissoon L, Ligon K, Wen PY, Bindra RS, Woo J, Arnaout O, Gerstner ER, Zhang PJ, Rosen BR, Yang L, Huang RY, Kalpathy-Cramer J. *Clin Cancer Res.* 2018 Mar 1;24(5):1073-1081.
- An introduction and overview of machine learning in neurosurgical care.  
**Senders JT**, Zaki MM, Karhade AV, Chang B, Gormley WB, Broekman ML, Smith TR, Arnaout O. *Acta Neurochir (Wien)*. 2018 Jan;160(1):29-38.
- Venous thromboembolism and intracranial hemorrhage after craniotomy for primary malignant brain tumors: a National Surgical Quality Improvement Program analysis.  
**Senders JT**, Goldhaber NH, Cote DJ, Muskens IS, Dawood HY, De Vos FYFL, Gormley WB, Smith TR, Broekman MLD. *J Neurooncol.* 2018 Jan;136(1):135-145.
- Machine learning and neurosurgical outcome prediction: a systematic review.  
**Senders JT**, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, Smith TR, Arnaout O. *World Neurosurg.* 2018 Jan;109:476-486.e1.
- Natural and artificial intelligence in neurosurgery: a systematic review.  
**Senders JT**, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, Smith TR. *Neurosurgery*. 2018 Aug 1;83(2):181-192.
- Innovation in neurosurgery: less than IDEAL? A systematic review.  
Muskens IS, Diederend SJH, **Senders JT**, Zamanipoor Najafabadi AH, van Furth WR, May AM, Smith TR, Bredenoord AL, Broekman MLD. *Acta Neurochir (Wien)*. 2017 Oct;159(10):1957-1966.

- Visual outcomes after endoscopic endonasal pituitary adenoma resection: a systematic review and meta-analysis.  
Muskens IS, Zamanipoor Najafabadi AH, Briceno V, Lamba N, **Senders JT**, van Furth WR, Verstegen MJT, Smith TRS, Mekary RA, Eenhorst CAE, Broekman MLD. *Pituitary*. 2017 Oct;20(5):539-552.
- Agents for fluorescence-guided glioma surgery: a systematic review of preclinical and clinical results.  
**Senders JT**, Muskens IS, Schnoor R, Karhade AV, Cote DJ, Smith TR, Broekman ML. *Acta Neurochir (Wien)*. 2017 Jan;159(1).
- The woven endobridge device for treatment of intracranial aneurysms: a systematic review.  
Muskens IS, **Senders JT**, Dasenbrock HH, Smith TR, Broekman ML. *World Neurosurg*. 2017 Feb;98:809-817.e1.
- Increased power of resting-state gamma oscillations in autism spectrum disorder detected by routine electroencephalography.  
van Diessen E, **Senders JT**, Jansen FE, Boersma M, Bruining H. *Eur Arch Psychiatry Clin Neurosci*. 2015 Sep;265(6):537-40.

## List of presentations

### *Invited presentations*

- The basics of artificial intelligence and machine learning and their applications in neurosurgery. **Senders JT**. *King's College London International Neurosurgical Conference*. 2018, London, UK.
- Applied machine learning in neurosurgical oncology. **Senders JT**. *Refereeravond Neuro-Oncologie*. 2018, The Hague, NL.
- Agents for fluorescence guided glioma surgery: a systematic review of preclinical and clinical results. **Senders JT**, Muskens IS, Schnoor R, Karhade AV, Cote DJ, Smith TR, Broekman MLD. *European Association of Neurological Surgeons' Annual Scientific Meeting*, 18<sup>th</sup> meeting.

*Other oral presentations*

- An online calculator for the prediction of survival for glioblastoma patients using classical statistics and machine learning. **Senders JT**, Staples P, Mehrtash A, Cote DJ, Gormley WB, Smith TR, Broekman MLD, Arnaout O. *Landelijke Werkgroep Neuro-Oncology's Annual Scientific Meeting. 2018, Utrecht, NL.*
- Venous thromboembolism and intracranial hemorrhage after craniotomy for primary malignant brain tumors: A National Surgical Quality Improvement Program analysis. **Senders JT**, Gaber N, Cote DJ, Muskens IS, Dawood HY, de Vos FYFL, Gormley WB, Smith TR, Broekman MLD. *European Association of Neurological Surgeons' (EANS) Annual Scientific Meeting, 17<sup>th</sup> meeting. 2017, Venice, IT.*
- Length of thromboprophylaxis in patients operated for a high-grade glioma: a retrospective study. **Senders JT**, Snijders TJ, Van Bentum GM, Seute T, De Vos FY, Robe PA, Broekman MLD. *Scientific day of the National Society of Neuro-Oncology. 2015, Utrecht, NL.*

## Courses

2018 – 2019	<b>Certificate in Applied Biostatistics (140h)</b> <i>Harvard Catalyst, Boston, MA, USA</i>
2018	<b>Introduction to Mixed Methods Research (32h)</b> <i>Harvard Catalyst, Boston, MA, USA</i>
2018	<b>Implementation Research (20h)</b> <i>Harvard Catalyst, Boston, MA, USA</i>
2016 – 2018	<b>Data Science, Machine Learning, Statistics in R and Python (156h)</b> <i>DataCamp, Cambridge, MA, USA</i>
2018	<b>Comparative Effectiveness Research (48h)</b> <i>Harvard Catalyst, Boston, MA, USA</i>
2017	<b>Statistics and R (16h)</b> <i>Harvard edX, Cambridge, MA, USA</i>
2017	<b>Leadership Strategies for the Researcher (16h)</b> <i>Harvard Catalyst, Boston, MA, USA</i>
2017	<b>Effectively Communicating Research (16h)</b> <i>Harvard Catalyst, Boston, MA, USA</i>
2016	<b>Meta-analysis Application (42h)</b> <i>MCPHS University, Boston, MA, USA</i>
2016	<b>Quantitative Methods (42h)</b> <i>MCPHS University, Boston, MA, USA</i>
2016	<b>Business Course (8h)</b> <i>Gupta Strategy Consultants, Rotterdam, NL</i>
2013	<b>Global Health Course (24h)</b> <i>University Medical Center Utrecht, Utrecht, NL</i>
2013	<b>Laparoscopic suturing course (9h)</b> <i>Leiden University Medical Center, Utrecht, NL</i>
2012	<b>Leadership and Teambuilding (17h)</b> <i>Birkbeck University, London, UK</i>



# CURRICULUM VITAE

Joeky Senders, born November 22, 1991 in Utrecht, graduated *summa cum laude* from Utrecht's Stedelijk Gymnasium in 2010, after which he started medical school at Utrecht University. Throughout his studies he completed two honors programs, was member of two student associations, and did multiple electives and rotations abroad including Birkbeck University (London, UK), 's Lands Hospitaal (Paramaribo, SR), Sanglah Hospital (Denpasar, ID), and Karapitiya Teaching Hospital (Galle, LK).

In July 2016, Joeky moved to Boston to work as a clinical research fellow under supervision of Dr. Marike Broekman and Dr. Timothy Smith at the Department of Neurosurgery in Brigham and Women's Hospital, Harvard Medical School. Throughout this two-year research fellowship, he investigated the application of machine learning in neurosurgical oncology, which resulted in the completion of the current thesis. Joeky was awarded with various research grants, has (co-)authored over 30 peer-reviewed publications, developed the Glioblastoma Survival Calculator and an open-source natural language processing framework for clinical text mining, worked as editor for 4Neurosurgery, and presented his work at several national and international conferences, recently receiving the "Best Publication in Neuro-Oncology Award" at the Annual Meeting of the European Association of Neurological Surgeons.

Joeky graduated *cum laude* from medical school at Utrecht University in August 2019 after which he started as neurosurgical resident not in training (ANIOS) at Haaglanden Medical Center in the Hague. In December 2019, he started as neurosurgical resident in training (AIOS) at University Medical Center in Utrecht.



