Big data, machine learning, and artificial intelligence: a field guide for neurosurgeons

Bharath Raju, MD, MCh,1 Fareed Jumah, MD,1 Omar Ashraf, BA,1 Vinayak Narayan, MD,1 Gaurav Gupta, MD,1 Hai Sun, MD, PhD,1 Patrick Hilden, MS,2 and Anil Nanda, MD, MPH1

1Department of Neurosurgery, Rutgers—Robert Wood Johnson Medical School and University Hospital; and 2Rutgers Neurosurgery Health Outcomes, Policy, and Economics (HOPE) Center, New Brunswick, New Jersey

Big data has transformed into a trend phrase in healthcare and neurosurgery, becoming a pervasive and inescapable phrase in everyday life. The upsurge in big data applications is a direct consequence of the drastic boom in information technology as well as the growing number of internet-connected devices called the Internet of Things in healthcare. Compared with business, marketing, and other sectors, healthcare applications are lagging due to a lack of technical knowledge among healthcare workers, technological limitations in acquiring and analyzing the data, and improper governance of healthcare big data. Despite these limitations, the medical literature is flooded with big data–related articles, and most of these are filled with abstruse terminologies such as machine learning, artificial intelligence, artificial neural network, and algorithm. Many of the recent articles are restricted to neurosurgical registries, creating a false impression that big data is synonymous with registries. Others advocate that the utilization of big data will be the panacea to all healthcare problems and research in the future. Without a proper understanding of these principles, it becomes easy to get lost without the ability to differentiate hype from reality. To that end, the authors give a brief narrative of big data analysis in neurosurgery and review its applications, limitations, and the challenges it presents for neurosurgeons and healthcare professionals naive to this field. Awareness of these basic concepts will allow neurosurgeons to understand the literature regarding big data, enabling them to make better decisions and deliver personalized care.

https://thejns.org/doi/abs/10.3171/2020.5.JNS201288

KEYWORDS big data; machine learning; artificial intelligence; robotics; artificial neural network; deep learning; algorithm

Gaining knowledge, is the first step to wisdom. Sharing data, is the first step to humanity.

— Unknown

The healthcare sector is witnessing a paradigm shift due to the recent growing trend of mass digitization of its information. Most hospitals in the world are utilizing electronic health records (EHRs) for recording all patient-related data,1 and, due to this digitization, a lot of structured and unstructured data is generated.2 Also, technological advancement has allowed data to be continuously generated from medical devices and wearable Internet of Things (IoT) devices. This has led to an exponential growth of healthcare data over a short period and resulted in what is known as big data.2–4 Despite having growth, there is a significant gap in the knowledge within the neurological community regarding big data and its implications. Although there are a few excellent reviews about big data in the neurological literature, most of them focus on neurosurgical registries, thereby creating a false impression in the mind of the reader that big data solely comprises big registries.5,6 Few go further, using specialized terminologies like artificial intelligence (AI), machine learning, artificial neural network (ANN), deep learning, and natural language processing without clearly defining these concepts,7,8 thus restricting the knowledge and use of this field to only a handful.

Despite having access to this vast amount of data, physicians and surgeons lack the time, tools, and expertise to take full advantage of what it has to offer.9 At the other extreme, some practitioners believe that the utilization of big data will provide solutions to all healthcare-related problems and research questions.10 To this end, we report on the capabilities as well as the limitations and challenges of big data from the time of data acquisition to the final visualization of the outcome. Besides, we define some of
the pertinent terminologies in this field and describe their relation to big data. Thus, we aim to equip neurosurgeons with the knowledge to differentiate between the hype and reality of big data to understand the role it will play in the future of neurosurgery and healthcare.

**Data Science**

Data science is a specialty in computer science. Kuhlen et al. noted, "The field of big data, machine learning, deep learning, and algorithm development and validation is often referred to as data science." Figure 1 depicts the overall relationship between data science and the different subdivisions or terminologies associated with it. As Drew Conway has stated, his Venn diagram visually illustrates that "data science is a mix of content expertise, methodological knowledge, and IT [information technology] skills" (Fig. 2). It shows that neither a computer scientist without neurosurgical expertise nor a neurosurgeon without computational knowledge can contribute to data science and its applications. As such, a proper collaboration between those with subject-specific expertise and technical knowledge is critical to make sense of the data generated.

**Big Data**

"Big data" is a term used to describe exceedingly large data sets, characterized by their complexity and variety. Big data is commonly represented by the 4 V’s: volume, variety, velocity, and veracity.

**Volume: So Large**

How big is big data? In fact, there is no clear-cut value to define big, as it is continually evolving, a direct consequence of a period’s technological capabilities. Kuo et al. concluded in their analysis that to be considered big data it should be at least in petabytes (2\(^{50}\) bytes, or 1024 terabytes). Healthcare data is now measured in zettabytes (2\(^{70}\) bytes, or a billion terabytes) or yottabytes (2\(^{80}\) bytes, or 1024 zettabytes).

**Variety: So Dispersed**

The variety in big data is due to different formats of data from various sources such as wearable devices to the hospital registry. Structured data is the data that can be easily stored, manipulated, or analyzed using a machine such as patient details in EHRs, whereas, unstructured data, for example, MRI, CT, and PET, cannot be queried or analyzed by a machine directly. Currently, 90% of big data is in the form of unstructured data (Table 1).
Velocity: Growing So Rapidly

Such volume and variety of data are continuously generated and added in a short time. Compared with relatively static data such as the EHR, the real-time streaming data from sources such as monitoring devices and imaging data further accelerate the velocity of growing data.13

Veracity: So Accurate

Veracity refers to the correctness and accuracy of the data obtained. To obtain a meaningful outcome from the data, it is an absolute necessity to check the accuracy of data and the quality of the source from which it is derived.13 Unfortunately, due to enormous volume and variety, coupled with the velocity at which it is obtained, data can never be entirely accurate, and hence it is challenging to validate.

Various authors have proposed the use of additional V’s to characterize big data, including variability, value, viscosity, volatility, and virality. We refer the interested reader to Fundamentals of Clinical Data Science by Kubben et al. for an in-depth review of these and other concepts.3,4,11 Before we proceed further to discuss the applications of big data, it is imperative to understand a few technical concepts and terminologies, such as algorithms and machine learning, that are used in data science.

Algorithm

“An algorithm ... is a set of heuristics and calculations that creates a model from data.”16 Essentially, it is a statistical or mathematical formula to process data. It can be used in data mining or machine learning. Data mining is the uncovering of hidden patterns and unknown information within the big data by a researcher (i.e., human learning). Here, we are clustering the data into intelligible pieces. Unlike data mining, machine learning models use data to improve a program’s or machine’s understanding of a hidden pattern.15

Machine Learning

Machine learning is the ability of a machine to learn from data. It is nothing but the creation of an algorithm. A machine or program can learn in many ways, such as supervised, unsupervised, semisupervised, and reinforcement learning, each comprising several algorithms (Table 2). There are many overlaps between each type of learning. Since a detailed discussion of each algorithm may confuse naive readers, here we give a general overview of each kind of learning and then a few commonly used algorithms or techniques. Though technical, understanding these concepts will aid in interpreting recent neurosurgical research using machine learning techniques.11,17

In supervised learning, also called predictive learning, a machine uses “completely labeled data” to learn using a general prediction rule to derive an algorithm. For example, one application may be in the prediction of survival of patients admitted to the ICU for traumatic brain injury. It starts with the creation of the training data set having details of traumatic brain injury patients, labeled alive or dead, admitted previously to the ICU. Machine learning algorithms such as ANNs are trained using this labeled data set. After the machine learns using this labeled data set, the algorithm will be validated by a distinct testing data set. After each future prediction, the model compares the output to the labeled data and adjusts itself when the projections are wrong, thus creating a final intelligent prediction model, which can be used on a new patient to predict their survival.18,19

In unsupervised learning, also called descriptive learning, there is no labeled or training data set. Here, the algorithm is fed with data without an explanation of what to do. The goal here is to find structure in the unknown data set by comparing the similarities within the data set.8,13 Lee et al. used an unsupervised machine learning algorithm to find the proportions of brain parenchyma, vasculature, and CSF within the nidus of brain arteriovenous malformations on T2-weighted MRI. Their model outperformed visual identification methods.20

Semisupervised learning is a balance between both supervised and unsupervised learning. In this approach, there are only a few labeled data in a larger pool of unlabeled data compared with completely labeled data in supervised learning. Consequently, the machine uses the labeled data to model an algorithm, which then uses the remaining unlabeled input data to extract output. Thus, the single database contributes to both testing and validation of the models. This method is useful when labeling data is difficult and time intensive. Using this technique, Scalzo and Hu created models that significantly reduced the rate of false alarms during intracranial pressure monitoring.21

In reinforcement learning, the algorithm tries to optimize a treatment outcome by a rewarding action taken in each step. Here, the algorithm is goal-oriented and used in sequential decision-making tasks. Data is exploited in each step to explore new ways of reaching the

### Table 1. Format and sources of big data

<table>
<thead>
<tr>
<th>Format</th>
<th>Primary Source</th>
<th>Secondary Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured: labeled data that is easily searchable</td>
<td>Hospital billing data; actuarial/insurance data; lab instruments reading; sensor data; coded EHR data: ICD code, CPT code, SNOMED CT, LOINC</td>
<td>Hospital; clinic; lab; physician offices; from everywhere due to IoT</td>
</tr>
<tr>
<td>Unstructured: textual or non textual unlabeled data that cannot be searchable</td>
<td>EHR patient notes; prescriptions/discharge summary; MRI/CT/x-ray; genetic data; physiological data: EEG, neuromonitoring, etc.; social media; mobile phones; intelligent wearable and fitness devices</td>
<td></td>
</tr>
<tr>
<td>Semistructured</td>
<td>Extensible markup language (XML) data</td>
<td></td>
</tr>
</tbody>
</table>

CPT = Current Procedural Terminology; LOINC = Logical Observation Identifiers Names and Codes; SNOMED CT = Systematized Nomenclature of Medicine–Clinical Terms.
desired goal. The action or path taken to achieve the desired goal is rewarded. This is repeated over many steps to optimize the decision-making sequence and get the long-term outcome. The problem with this kind of learning is that the end results after multiple stepwise decisions are difficult to predict at the beginning.22,23 Reinforcement learning has been used in the customization of antiretroviral and antiepileptic drug therapy, as well as in sepsis management.23 Though this technique looks promising, its application in neurosurgery is still underutilized and evolving.

In general, the type of learning used is based on the characteristics of the input data and whether the outcome variable is predefined or not. If the data set contains a known variable, then supervised learning is preferred. Within supervised learning, different algorithms are used based on whether the predicted outcome is a discrete or a continuous variable. For example, a regression algorithm is used if the outcome variable is continuous, and classification algorithms if it is discrete. An unsupervised algorithm such as clustering and dimension reduction is used to describe an unknown pattern from a data set. Other factors to consider are the quality of the data set, the type of input variable (image/text), and the relationship between the input and output variables (linear or nonlinear), among others.24

**ANN and Deep Learning**

ANN and deep learning are the most referenced algorithms in the neurosurgical literature. There is no clear-cut definition for ANN. In simplest terms, the ANN is comparable to a biological neural network, where training an ANN is akin to learning in the brain. ANNs are characterized by input and output layers (comprising input and output “neurons” or nodes/units) with hidden layers in between. These nodes store different information, and they interact with other nodes to arrive at the best possible outcome. The hidden neurons can vary from a few to thousands. If this hidden layer of a neuron contains multiple layers, it is called deep learning. Increasing the number of hidden layers to a certain extent improves training and modeling at the expense of more demanding computational requirements (Fig. 3).11,17 So, developing a model with the right number of layers is of utmost importance.

**Ensemble**

Many researchers use the ensemble machine learning technique, wherein they combine multiple models or algorithms to improve the overall accuracy of the predictions. There are many methods of ensemble learning, such as bagging (random forest model) and boosting, among others. Without going into technical details, it is crucial to understand that these techniques reduce the noise, variability, and bias induced by each model individually.25 Most of the research in neurosurgical journals, especially involving radiological data, uses this technique. Table 3 shows many recently published machine learning–related neurosurgical articles in which many of the methods mentioned above have been used. After gaining some insights into these concepts, it is pivotal to get the hang of the technology behind big data analysis to understand the complexities and need for technical expertise.

**Big Data Technology**

Usual scientific research on any data is done by statistical analysis of data to arrive at an outcome. Data is entered as a table on your computer hard disc and analyzed using a statistical methodology for reasoning the association between the data and outcome. Due to the volume and complexity of data from different sources, various technologies are used to arrive at a result.

The hardware used to store the data are called the platforms. It can be local storage or a distributed storage system. In a local storage system, data is stored in one place,
usually at the source of data. In a distributed storage system, also known as a cloud platform, data is distributed between multiple systems or computers at different locations. Most healthcare systems use cloud platforms due to cost-effectiveness and availability. Hadoop Distributed File Systems, Amazon S3, Google Cloud Services, and Microsoft Azure are the most used platforms. Frameworks are the interfaces to analyze the data stored on platforms; Hadoop and Spark are widely used frameworks. Figure 4 shows the comparison of big data analytics with conventional research methodology, highlighting the differences at each stage of the data analysis workflow.

**Why Do We Need Such Big Data?**

Knowing the relationship between big data, machine learning, and AI is a prerequisite to understanding the importance of data and registries in neurosurgery. Most of the big data applications are directly or indirectly related to the field of data analytics, AI, and robotics (Fig. 5). Here, we give a brief insight into AI and robotics, followed by more general applications of big data in the medical field.

**AI and Robotics**

In general, AI refers to a subfield of data science and computer science, in which computers perform tasks that would typically require human intelligence. The term “artificial intelligence” often conjures up images of robots, though this is generally not the case; robotics is just one component of AI. It is giving the machines the ability to sense, reason, engage, and learn. Robotics and motion, natural language processing, voice recognition, and computer vision are types of AI. AI combines and uses machine learning algorithms, ANN, deep learning, and other data analytical techniques to gain intelligent capabilities. Recently, the US FDA has cleared a deep-learning algorithm-based AI called Viz.ai that analyzes images to detect potential large-vessel occlusion strokes on CT angiography and alert the specialist to initiate emergency treatment, whereas in robotics, machines with mechanical parts use the same methods of learning to think and perform like humans.

**Clinical Outcome Analysis**

Machine learning can be used to improve diagnostics as well as in the prediction of patient outcomes and diagnosis. Many studies have proven that predictions derived from machine learning may be better than conventional statistical models of prediction when applied to neurological patients. A few recently published studies in leading neurosurgical journals and the types of questions answered by those studies are listed in Table 3.

**Personalized/Precision Medicine in Neurosurgery**

Personalized and precision medicine involves tailoring the medical or surgical treatment to the characteristics of an individual patient. Instead of subgrouping patients based on only phenotypic characteristics, it uses imaging, genomics, proteomics, and epigenetic information in an effort to individualize treatment options and improve prognostication. Multiple data points extracted from each of these sources, along with EHR data, constitute the big data. These big data are analyzed using machine learning and AI to diagnose, prognosticate, and allocate individualized treatment, thereby reducing unnecessary drug use, toxicity, and cost, as well as optimizing treatment. Glioblastoma is a prototypical example in neurosurgery where each tumor can be grouped and subgrouped.
<table>
<thead>
<tr>
<th>Article Title</th>
<th>Authors &amp; Year</th>
<th>Journal</th>
<th>Algorithms Used in the Study</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicting inpatient length of stay after brain tumor surgery: developing</td>
<td>Muhlestein et</td>
<td>Neurosurg</td>
<td>7 machine learning algorithms trained: RF, XGBoost, GLMs, BDT,</td>
<td>Predictive</td>
</tr>
<tr>
<td>machine learning ensembles to improve predictive performance</td>
<td>et al., 2019</td>
<td></td>
<td>K-NN, GLMs, ANN</td>
<td></td>
</tr>
<tr>
<td>An online calculator for the prediction of survival in glioblastoma patients</td>
<td>Senders et al.,</td>
<td>Neurosurg</td>
<td>15 machine learning algorithms trained</td>
<td>Predictive</td>
</tr>
<tr>
<td>using classical statistics and machine learning</td>
<td>2020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A traumatic brain injury prognostic model to support in-hospital tri-</td>
<td>Hernandes</td>
<td>J Neurosurg</td>
<td>9 machine learning algorithms trained: K-NN, Bayesian GLM, etc.</td>
<td>Prognostic</td>
</tr>
<tr>
<td>age in a low-income country</td>
<td>Rocha et al.,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural network–based identification of patients at high risk for intra-</td>
<td>Staatjes et al.,</td>
<td>J Neurosurg</td>
<td>DL</td>
<td>Predictive</td>
</tr>
<tr>
<td>operative cerebrospinal fluid leaks in endoscopic pituitary surgery</td>
<td>2020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning applied to neuroimaging for diagnosis of adult classic</td>
<td>Urbizu et al.,</td>
<td>J Neurosurg</td>
<td>7 machine learning algorithms trained: NB, DT, K-NN, LR, SVM,</td>
<td>Diagnostic</td>
</tr>
<tr>
<td>Chiari malformation: role of the basion as a key morphometric indicator</td>
<td>2018</td>
<td></td>
<td>LDA</td>
<td></td>
</tr>
<tr>
<td>Using artificial neural networks to identify patients with concussion and</td>
<td>Landry et al.,</td>
<td>J Neurosurg</td>
<td>ANN</td>
<td>Diagnostic</td>
</tr>
<tr>
<td>postconcussion syndrome based on antiscadacdes</td>
<td>2019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A computer vision approach to identifying the manufacturer and model of</td>
<td>Huang et al.,</td>
<td>J Neurosurg</td>
<td>KAZE feature extractor, K-means clustering, SVM</td>
<td>Computer vision</td>
</tr>
<tr>
<td>anterior cervical spinal hardware</td>
<td>2019</td>
<td>Spine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using machine learning to predict 30-day readmissions after posterior</td>
<td>Hopkins et al.,</td>
<td>J Neurosurg</td>
<td>DL</td>
<td>Predictive</td>
</tr>
<tr>
<td>lumbar fusion: an NSQIP study involving 23,264 patients</td>
<td>2020</td>
<td>Spine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning for automated 3-dimensional segmentation of the spine and</td>
<td>Burström et al.,</td>
<td>J Neurosurg</td>
<td>Multiple segmentation algorithms trained (not mentioned)</td>
<td>Predictive</td>
</tr>
<tr>
<td>suggested placement of pedicle screws based on intraoperative cone-beam</td>
<td>2019</td>
<td>Spine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>computer tomography</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can machine learning algorithms accurately predict discharge to nonhome</td>
<td>Goyal et al.,</td>
<td>J Neurosurg</td>
<td>7 machine learning algorithms trained: predictive hierarchical</td>
<td>Predictive</td>
</tr>
<tr>
<td>facility and early unplanned readmissions following spinal fusion? Analysis</td>
<td>2019</td>
<td>Spine</td>
<td>clustering, classification algorithm</td>
<td></td>
</tr>
<tr>
<td>of a national surgical registry</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning modeling for predicting hospital readmission following</td>
<td>Kalagara et al.,</td>
<td>J Neurosurg</td>
<td>CT</td>
<td>Predictive</td>
</tr>
<tr>
<td>lumbar laminectomy</td>
<td>2018</td>
<td>Spine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning–based preoperative predictive analytics for lumbar</td>
<td>Siccoli et al.,</td>
<td>Neurosurg</td>
<td>7 machine learning algorithms trained: RF, XGBoost, GLMs, BDT,</td>
<td>Predictive</td>
</tr>
<tr>
<td>spinal stenosis</td>
<td>2019</td>
<td>Focus</td>
<td>K-NN, GLMs, ANN</td>
<td></td>
</tr>
<tr>
<td>Machine learning applications for the prediction of surgical site</td>
<td>Tunthanathip et</td>
<td>Neurosurg</td>
<td>DT, NB w/ Laplace correction, K-NN, ANN</td>
<td>Predictive</td>
</tr>
<tr>
<td>infection in neurological operations</td>
<td>et al., 2019</td>
<td>Focus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development of machine learning algorithms for prediction of discharge</td>
<td>Karhade et al.,</td>
<td>Neurosurg</td>
<td>RF, ANN, SVM, Bayes point machine, BDT</td>
<td>Predictive</td>
</tr>
<tr>
<td>disposition after elective inpatient surgery for lumbar degenerative disc</td>
<td>2018</td>
<td>Focus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome prediction of intracranial aneurysm treatment by flow diverters</td>
<td>Paliwal et al.,</td>
<td>Neurosurg</td>
<td>SVM, LR, K-NN, ANN</td>
<td>Predictive</td>
</tr>
<tr>
<td>using machine learning</td>
<td>2018</td>
<td>Focus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A machine learning approach to predict early outcomes after pituitary</td>
<td>Hollon et al.,</td>
<td>Neurosurg</td>
<td>LR w/ elastic net</td>
<td>Predictive</td>
</tr>
<tr>
<td>adenoma surgery</td>
<td>2018</td>
<td>Focus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning analyses can differentiate meningioma grade by features</td>
<td>Hale et al.,</td>
<td>Neurosurg</td>
<td>K-NN, SVM, NB, ANN</td>
<td>Predictive</td>
</tr>
<tr>
<td>on magnetic resonance imaging</td>
<td>2018</td>
<td>Focus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utility of deep neural networks in predicting gross-total resection after</td>
<td>Staartjes et al.,</td>
<td>Neurosurg</td>
<td>DL</td>
<td>Predictive</td>
</tr>
<tr>
<td>transtemporal surgery for pituitary adenoma: a pilot study</td>
<td>2016</td>
<td>Focus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development and validation of an automatic segmentation algorithm for</td>
<td>Scherer et al.,</td>
<td>Stroke</td>
<td>RF</td>
<td>Predictive</td>
</tr>
<tr>
<td>quantification of intracerebral hemorrhage</td>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction of IDH1 mutation status in glioblastoma using machine learning</td>
<td>Lee et al.,</td>
<td>World Neu-</td>
<td>Classification algorithms: K-NN, SVM, DT, RF, NB, LDA, GBM</td>
<td>Predictive</td>
</tr>
<tr>
<td>technique based on quantitative radiomic data</td>
<td>2019</td>
<td>rosurg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BDT = boosted DT; DL = deep learning; DT = decision tree; GBM = gradient boosting machine; GLM = generalized linear model; K-NN = K-nearest neighbors; LDA = linear discriminant assay; LR = logistic regression; NB = naive Bayes classifier; RF = random forest; SVM = support vector machine; XGBoost = extreme gradient boosting.
based on classical histopathology, oncogenic events, and signaling pathways. Individualized treatment can be delivered by targeting the specific gene or molecules affected in those pathways. 49

Medical Device Design and Manufacturing
The prospects of AI in neurotechnology and bioelectronic medicine are exciting. Numerous brain-computer interfaces are used to pick up neurological signals, such

**FIG. 4.** Comparison between conventional research and big data research. HDFS = Hadoop Distributed File System; HQL = Hibernate Query Language; SQL = Structured Query Language; VBA = Visual Basic for Applications. Figure is available in color online only.

**FIG. 5.** Relationship between big data, machine learning, AI, robotics, and data analytics. A: Data analytical pathway where a machine learning (ML) model is used to predict, prognosticate, or diagnose. B: The red dotted line shows the robotics pathway. C: The blue dotted line shows the AI pathway. Machine learning models created by big data analytics are central to all 3 pathways. NLP = natural learning process. Figure is available in color online only.
as EEG, which can be transmitted to a computer where these signals are decoded and sent to effector organs such as limb muscles or peripheral nerves to cause the desired movement. Through machine learning, algorithms are trained to make sense of those neurological signals and respond accordingly to optimize outcomes. One example includes closed-loop deep brain stimulation, in which electrodes are implanted in the desired location of the CNS to modulate the electrical activity in order to control function. AI can be utilized to monitor vital signs or other electrophysiological indicators (motor evoked potentials) to provide feedback and improve modulation control.\textsuperscript{50,51}

**Drug Development**

AI can analyze drug molecules and predict their functionality. AI-based virtual development tools can enable the development of a new molecule or the identification of a new molecular target in a virtual environment, from which the best molecule can be selected for clinical trials. One such example includes AI IBM Watson, which was used in a study to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis.\textsuperscript{52} AI can also identify patients who might respond better to a certain medication and predict possible complications, thus optimizing patient selection, augmenting the speed of discovery, and reducing the cost burden.\textsuperscript{3,53} Clinical trial endpoints and drug compliance can also be determined using AI-based technologies.\textsuperscript{54}

**Medical Education and Decision-Making**

Processing and organizing medical information is sometimes beyond human capabilities and may warrant the use of intelligent machines. AI can be used in knowledge acquisition and knowledge representation. This allows the development of reasoning tools to help with the decision-making process. Many cognitive apprenticeship models have been developed for this purpose. Projects such as RadTutor,\textsuperscript{55} InforMed, SAFARI, the cardiac tutor, and GUIDON are a few examples.\textsuperscript{56}

**Managed Care**

The predominant fee-for-service system had been changed to a pay-for-performance system following the Affordable Healthcare Act in 2010. Both the public and private healthcare policymakers reward doctors and the hospital for improving the quality of care. Big data serves as the bedrock for the measurement of quality of care to reward a healthcare provider. Health plans use big data to score and rank providers, which in turn is used to reward or withhold payment. Big data is also used for risk stratification of patients. This enables providers to identify high-risk patients, anticipate high-risk behavior, and implement targeted care management strategies individualized for each patient.\textsuperscript{57}

**Preventive Care**

IoT devices, such as the Apple Watch and Fitbit, can generate continuous and real-time data. These are components of big data within healthcare that can be influenced by AI. The use of AI in the analysis of this information has significant clinical utility in both primary and secondary prevention (i.e., disease prevention and early-stage intervention). For example, the present COVID-19 outbreak is modeled using big data to estimate the number of cases and deaths both with and without social distancing and based on the fact that many countries, including the US, imposed preventive measures such as tracking, quarantine, and evaluation to control the pandemic to a certain extent.\textsuperscript{58,59}

**Limitations**

**Data Acquisition, Storage, and Analysis**

Data is being continuously produced at an exponential rate. Traffic and volume have emerged as the two main issues in making use of this ever-increasing big data.\textsuperscript{13} One way to address this issue is by utilizing either platforms with extensive storage or clusters of computers. For physicians, simple and rapid retrieval of patient information to derive a solution is a necessity in day-to-day patient care. While the utilization of big data will give providers the ability to answer more complex questions to improve patient care, this increased complexity requires more time and analytical power. These computationally intense problems require expensive supercomputers with iterative computational power and real-time processing capabilities. The parallelization of computing or algorithms are needed for efficient and timely computing. Maintaining such an infrastructure is not currently cost-effective for many institutions, causing inadequate usage of available health data. Even beyond the issue of infrastructure, variability in the level of knowledge in healthcare and data management training leads to multitudes of errors in data entry and writing, precluding proper use of big data for scientific progress.\textsuperscript{13}

**Data Organization and Integration**

Healthcare data is complex and noisy due to the multiple sources and modalities that contribute to its development. Noise in the quality of data is often a consequence of missing values and unwanted data. As the quality of the data decreases, so too does the quality of the information extrapolated from that data decrease. As such, tools to allow data cleansing have been developed to improve “noisy” data, consequently enhancing the quality of information as well as the decision-making process.\textsuperscript{13}

Another issue involves the inability to properly integrate healthcare data. In an ideal scenario, standardized healthcare data would allow for seamless integration between databases, enabling optimal data compilation and subsequent analysis. In reality, there is a lack of standardization of healthcare data, introducing a source of error into this process. This becomes evident when looking at the functional differences between database software. For example, Gamma Knife machine data cannot be combined with the EHR. Semantic differences exist in the metadata such that “sex” in one database and “gender” in another prevent proper integration. Furthermore, mapping differences in metadata also exist, such as using “patient name” in one system as opposed to “first name” and “last name” in another. Variations in the use of medical terminologies,
like “epilepsy” or “seizures” or “fits,” demonstrate another barrier to data integration.13

Data Ownership and Exchange

Hospitals own the data collected that will be used for quality control, insurance, patient care, and research purposes. However, many healthcare providers are hesitant to share this data when considering the market competition, thus precluding the meaningful use of collected big data. Beyond this, other barriers for data exchange exist, such as legal, network, and privacy/security issues.3,12 In terms of legal restrictions, many data governance acts such as HIPAA (Health Insurance Portability and Accountability Act), HITECH (Health Information Technology for Economic and Clinical Health Act), GINA (Genetic Information Nondiscrimination Act), and FOIA (Freedom of Information Act) exist to monitor privacy and security, usability, and accessibility of data.4 Sharing large amounts of data also imposes a significant networking challenge, requiring a faster and broader network range and storage capacity. However, the advent of cloud computing has made data sharing between two healthcare systems cost-effective and facile.13

Data privacy deals with the control of personal information and the policies implemented to prevent the identification of a patient. Multiple privacy-preserving methods have been introduced, such as de-identification, hybrid execution models (HybrEx), and identity-based anonymization.60 This contrasts with data security, which deals with the protection of data collected from misuse or malicious attacks. Cybersecurity breaches of the protected health information have become a major concern following the adoption of EHR to store patient information. With the focus on implementation of mandatory EHR systems, there appears to be a serious lack of security measures and adequate personnel training, leading to the compromise of millions of patients’ records in the US. A 2018 report showed around 5 million electronic medical record–related breaches and 130 million hacking-related breaches, accounting for about 85% of all breaches.61 The IBM 2019 data breach reports an 80% increase in breaches from the last report, and healthcare data loss accounts for a major share of $2 trillion cost due to cybercrime.62 In response to these serious attacks, hospitals have employed the use of various sophisticated techniques to protect patient data, including authentication, encryption, data masking, and access control.

Other challenges encountered during big data management are briefly categorized as interoperability issues, data mining issues, data visualization issues, technical and technological issues, and data overfitting.

Conclusions

The importance of big data and its applications is gaining momentum in the medical field and neurosurgery. A basic understanding of big data and its applications, such as machine learning and AI, as well as its importance in decision-making during patient care, is a necessity for all medical professionals. Despite its ability to optimize patient care, big data has several limitations and challenges, and a thorough understanding of these will allow clinicians to distinguish between hype and reality in this rapidly evolving field. That being said, technologies to overcome the existing limitations are evolving rapidly. Ultimately, it is our responsibility to keep pace with this evolution to derive a meaningful outcome and personalize management for each patient.

References


Disclosures
The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.