

Aug 01, 2025

Version 1

Publicly available datasets for artificial intelligence in neurosurgery: a systematic review V.1

DOI

<https://dx.doi.org/10.17504/protocols.io.n92ld6z67g5b/v1>

Bianca Chan¹, Brandon Kim¹, Ethan Schonfeld^{1,2}, George Nageeb^{1,2}, Aaradhya Pant^{1,2}, Adam Sjöholm^{1,2}, Ravi Medikonda³, Ummey Hani¹, Anand Veeravagu^{1,3}

¹Stanford Neurosurgical Artificial Intelligence and Machine Learning Laboratory, Stanford School of Medicine, Stanford University, Stanford, CA;

²Stanford University School of Medicine, Stanford University, Stanford, CA;

³Department of Neurosurgery, Stanford University School of Medicine, Stanford University, Stanford, CA

Veeravagu Lab



Bianca Chan

Stanford University

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.n92ld6z67g5b/v1>



Protocol Citation: Bianca Chan, Brandon Kim, Ethan Schonfeld, George Nageeb, Aaradhya Pant, Adam Sjöholm, Ravi Medikonda, Ummey Hani, Anand Veeravagu 2025. Publicly available datasets for artificial intelligence in neurosurgery: a systematic review. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.n92ld6z67g5b/v1>

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: August 01, 2025

Last Modified: August 01, 2025

Protocol Integer ID: 223883

Keywords: available neurosurgical datasets suitable for machine learning, available neurosurgical dataset, neurosurgery, available datasets for artificial intelligence, available dataset, dataset characteristic, machine learning, pubmed search

Abstract

We conducted a systematic review according to PRISMA guidelines to identify publicly available neurosurgical datasets suitable for machine learning. A PubMed search on February 8, 2025, yielded 267 articles, of which 86 met inclusion criteria. Each study was reviewed to extract dataset characteristics, model development details, validation status, availability, and citation impact.

Search Strategy

- 1 Perform a comprehensive literature search in the PubMed database using the following exact query: data release[Title/Abstract] OR "novel data"[Title/Abstract] OR "primary dataset"[Title/Abstract] OR "dataset"[Title/Abstract]) AND ("machine learning"[Title/Abstract] AND "artificial intelligence"[Title/Abstract] OR "AI"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "ML"[Title/Abstract] OR "Neural Networks"[Title/Abstract]) AND ("neurosurger*"[MeSH Terms] OR "neurosurgical"[Title/Abstract] OR "vascular neurosurgery"[Title/Abstract] OR "neurooncology"[Title/Abstract] OR "functional neurosurgery"[Title/Abstract] OR "spine"[Title/Abstract] OR "TBI"[Title/Abstract] OR "neurosurgical care"[Title/Abstract]) AND ("2016/01/01"[Date - Publication] : "3000"[Date - Publication]).

Screening

- 2 Manually and individually review all resulting studies from the search strategy by two researchers in parallel.
- 3 Apply the following inclusion criteria: (1) publication from 01/01/2016 and onward, (2) release of novel data (not a secondary analysis or analysis of previously released data), (3) already publicly available data or can apply for access, (4) data derived from human patients with neurosurgical relevant diagnosis in pre, peri, post-operative care or laboratory samples.
- 4 Apply the following exclusion criteria: (1) non-English articles, (2) non-primary research, (3) less than 100 data items, (4) non-peer reviewed articles and abstracts. Exclude studies that included <100 patients but 100 or more data items were not excluded due to the increase in few shot learning techniques available. The 100 data item requirement aims to exclude case series that focus on dedicated projects of dataset development. Furthermore, novel data must comprise the majority ($\geq 50\%$) of all data used for training for studies that use both private and public datasets. If the split is unclear, exclude the study. If data was cited in a previously published study, exclude the article unless access can be requested from the authors and the article is included. This assumption is included to avoid missing subdomains of data that the convention may be to request data directly from authors in the subdomain.
- 5 After articles are independently selected by the screeners based on the above criteria and assumptions, resolve any disagreements between the two screeners through discussion.

Data Extraction and Analysis

- 6 Obtain basic identifying information from each selected article including title, publication year, and PubMed ID.
- 7 To study generalizability and validation of the data/models, extract the following variables from the selected articles by two researchers in parallel: whether the data is labeled, whether the data source is multi-institutional, number of institutions, whether there are institutions that are non-tertiary academic centers, whether a trained model is included, whether the model is externally validated, and whether the model is publicly available (e.g., if code and weights are provided) and has a public application (e.g., available web application). To meet the criteria for external validation, ensure the model is evaluated with data from more than one institution.
- 8 To study model functionality and impact, collect the primary data type, the sample size of the primary data, the label type, the inference class if the paper included a baseline model, performance metrics of the model, and the number of article citations.
- 9 Resolve any disagreements between the two authors on the information from object extraction through discussion.
- 10 Prior to performing analysis, classify the primary data type into seven categories (i.e.: X-ray, CT, MRI, PET, clinical, sensor, video). Categorize the label type into seven categories (i.e.: diagnosis, segmentation, detection/localization, intervention, grading, image, outcome). Across all studies, classify the inference class of the baseline model into eight categories (i.e.: diagnosis, detection/localization, segmentation, intervention, measurement, grading, outcome, generative). Classify all baseline models into linear, convolutional neural network (CNN), segmentation, non-transformer NLP, transformer, and generative adversarial network (GAN).

Quality Assessment

- 11 To assess the methodological quality and risk of bias of included studies, develop a custom checklist reflecting critical aspects of dataset and model robustness. Evaluate each study for inclusion of multi-institutional data, use of external validation, public availability of code, presence of a public-facing application, use of task-appropriate performance metrics, and inclusion of outcome labels.
- 12 Assign a score from 0 to 6 based on the number of criteria met. Use this simple quality score to gauge the risk of bias and generalizability of each dataset.