Data in Brief 44 (2022) 108542

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset on acute stroke risk stratification from CT angiographic radiomics

Emily W. Avery[a], Jonas Behland[a,b], Adrian Mak[a,b], Stefan P. Haider[a,c], Tal Zeevi[a], Pina C. Sanelli[d], Christopher G. Filippi[e], Ajay Malhotra[a], Charles C. Matouk[f], Christoph J. Griessenauer[g,h,i], Ramin Zand[j], Philipp Hendrix[g,k], Vida Abedi[l,m], Guido J. Falcone[n], Nils Petersen[n], Lauren H. Sansing[o], Kevin N. Sheth[n], Seyedmehdi Payabvash[a,*]

[a] *Section of Neuroradiology, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar St, New Haven, CT 06510, USA*
[b] *CLAIM - Charité Lab for Artificial Intelligence in Medicine, Charité Universitätsmedizin Berlin, Charitépl.1, Berlin 10117, Germany*
[c] *Department of Otorhinolaryngology, University Hospital of Ludwig Maximilians Universität München, Ziemssenstraße 1, München 80336, Germany*
[d] *Section of Neuroradiology, Department of Radiology, Northwell Health, 300 Community Dr, Manhasset, NY 11030, USA*
[e] *Section of Neuroradiology, Department of Radiology, Tufts School of Medicine, 1 Washington St, Boston, MA 02111, USA*
[f] *Division of Neurovascular Surgery, Department of Neurosurgery, Yale University School of Medicine, 333 Cedar St, New Haven, CT 06510, USA*
[g] *Department of Neurosurgery, Geisinger Medical Center, 100N Academy Ave, Danville, PA 17822, USA*
[h] *Research Institute of Neurointervention, Paracelsus Medical University, Strubergasse 21, Salzburg 5020, Austria*
[i] *Department of Neurosurgery, Paracelsus Medical University, Strubergasse 21, Salzburg 5020, Austria*
[j] *Department of Neurology, Geisinger Medical Center, 100N Academy Ave, Danville, PA 17822, USA*
[k] *Department of Neurosurgery, Saarland University Medical Center, Kirrberger Str 100, Homburg 66421, Germany*
[l] *Department of Molecular and Functional Genomics, Geisinger Medical Center, 100N Academy Ave, Danville, PA 17822, USA*
[m] *Biocomplexity Institute, Virginia Tech, 1015 Life Science Cir, Blacksburg, VA 24061, USA*
[n] *Division of Neurocritical Care and Emergency Neurology, Department of Neurology, Yale University School of Medicine, 333 Cedar St, New Haven, CT 06510, USA*
[o] *Division of Stroke and Vascular Neurology, Department of Neurology, Yale University School of Medicine, 333 Cedar St, New Haven, CT 06510, USA*

* Corresponding author.
E-mail address: sam.payabvash@yale.edu (S. Payabvash).
Social media: @emilywavery (E.W. Avery), @AnotherMak (A. Mak), @sairaallapeikko (C.G. Filippi), @AjayMalhotraRad (A. Malhotra), @MatoukCharles (C.C. Matouk), @cgriessenauer (C.J. Griessenauer), @GuidoFalconeMD (G.J. Falcone), @LaurenHSansing (L.H. Sansing), @sheth_kevin (K.N. Sheth), @SamPayabvash (S. Payabvash)

https://doi.org/10.1016/j.dib.2022.108542

ARTICLE INFO

Article history:
Received 15 June 2022
Revised 2 August 2022
Accepted 10 August 2022
Available online 14 August 2022

Dataset link:
Dataset on acute stroke risk stratification from CT angiographic radiomics (Original data)

Keywords:
Machine-learning
Radiomics
Large vessel occlusion
Stroke
Telestroke
CTA

ABSTRACT

With advances in high-throughput image processing technologies and increasing availability of medical mega-data, the growing field of radiomics opened the door for quantitative analysis of medical images for prediction of clinically relevant information. One clinical area in which radiomics have proven useful is stroke neuroimaging, where rapid treatment triage is vital for patient outcomes and automated decision assistance tools have potential for significant clinical impact. Recent research, for example, has applied radiomics features extracted from CT angiography (CTA) images and a machine learning framework to facilitate risk-stratification in acute stroke. We here provide methodological guidelines and radiomics data supporting the referenced article "CT angiographic radiomics signature for risk-stratification in anterior large vessel occlusion stroke." The data were extracted from the stroke center registry at Yale New Haven Hospital between 1/1/2014 and 10/31/2020; and Geisinger Medical Center between 1/1/2016 and 12/31/2019. It includes detailed radiomics features of the anterior circulation territories on admission CTA scans in stroke patients with large vessel occlusion stroke who underwent thrombectomy. We also provide the methodological details of the analysis framework utilized for training, optimization, validation and external testing of the machine learning and feature selection algorithms. With the goal of advancing the feasibility and quality of radiomics-based analyses to improve patient care within and beyond the field of stroke, the provided data and methodological support can serve as a baseline for future studies applying radiomics algorithms to machine-learning frameworks, and allow for analysis and utilization of radiomics features extracted in this study.

Specifications Table

Subject	Medical Imaging
Specific subject area	Radiomics-based risk stratification in acute large vessel occlusion triage
Type of data	Table Figure Text
How the data were acquired	The data were acquired by retrospective electronic health record review at two institutions: Yale New Haven Hospital and Geisinger Medical Center. Patients in the Yale Stroke Center registry who presented between 1/1/2014–10/31/2020 and patients in the Geisinger Stroke Registry who presented between 1/1/2016–12/31/2019 were identified and included in the dataset based on clinical and imaging data availability.
Data format	Raw Analyzed
Description of data collection	Patients were included if they: (1) suffered anterior circulation large vessel occlusion (LVO), (2) underwent mechanical thrombectomy, (3) had CTA source images with slices $\leq$1 mm, and (4) had modified Rankin Scale (mRS)

(*continued on next page*)

	assessment of functional outcome recorded at discharge or 3-mo follow-up. Radiomics features were extracted from the anterior circulation territory of each admission CTA using FSL and pyRadiomics software.
Data source location	• Institution 1: Yale New Haven Hospital • City/Town/Region: New Haven, CT • Country: USA • Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: 41°18′14.7″N 72°56′07.0″W • Institution 2: Geisinger Medical Center • City/Town/Region: Danville, PA • Country: USA • Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: 40°58′04.0″N 76°36′17.7″W
Data accessibility	The referenced data is included as supplemental material in the submission, and is also available at our Github repository: https://github.com/emilywavery/Radiomics-data-sharing/tree/radiomicsdata
Related research article	Avery, E.W., Behland, J., Mak, A., Haider, S.P., Zeevi, T., Sanelli, P.C., Filippi, C.G., Petersen, N.H., Falcone, G.J., Sansing, L.H., Malhotra, A., Greissenauer, C.J., Zand, R., Hendrix, P., Abedi, V., Matouk, C.C., Sheth, K.N., Payabvash, S. CT angiographic radiomics signature for risk-stratification in anterior large vessel occlusion stroke. Neuroimage: Clinical, 2022;34:103034 [1]

Value of the Data

- The data included in this publication enrich the body of publicly available radiomics data, a field of growing interest in biomedical imaging research.
- The radiomics data included in this publication can be utilized in conjunction with the methodological guide of the related research article to serve as a point of comparison for researchers utilizing similar machine learning methodologies.
- The data can benefit researchers and clinicians interested in neuroimaging, stroke, and endovascular mechanical thrombectomy. It is also of use to researchers interested in radiomics, machine-learning, and artificial intelligence.
- Further insights and analyses that research groups may explore from our data include: support in radiomics-based analyses, comparison of radiomics features of this dataset to those of other datasets, and assessment of clinical and radiomics variables affecting stroke patient outcomes.

1. Data Description

The data files that appear in this article include:

(1) AnalyzedData.docx: Table 1 summarizes the machine learning and feature selection methods utilized in the related research article. Table 2A and 2B describe the clinical and demographic characteristics of patients from each study center.

(2) Radiomics_*.csv files: These files provide the values of all extracted radiomics features for the Yale and Geisinger datasets described in the reference article. These radiomics features were extracted from the bilateral middle cerebral artery (MCA) territories of each patient's admission CTA. A complete list of the first-order and texture features used in this study is described in van Griethuysen et al. [2], and exact feature definitions are described in Pyradiomics documentation [3]. Select first-order and texture features are also described in the related research article Supplementary Table 1 [1].
A separate file is provided for discharge (short-term) and 3-month (long-term) outcome cohorts for the Yale training/cross-validation (CV) dataset, independent Yale dataset, and external Geisinger dataset (3-month – long-term – outcome cohort only). The files are titled accordingly and include:
Radiomics _YaleTrainingCV_ShortTermFollowUP.csv

Radiomics _YaleTrainingCV_LongTermFollowUP.csv
Radiomics _YaleIndependent_ShortTermFollowUP.csv
Radiomics _YaleIndependent_LongTermFollowUP.csv
Radiomics _Geisinger_LongTermFollowUP.csv

(3) ClinicalData_*.csv files: These files provide the sex and age of each patient. A separate file is provided for discharge(short-term) and 3-month (long-term) outcome cohorts for the Yale training/CV dataset, independent Yale dataset, and external Geisinger dataset (long-term outcome cohort only). The files are titled accordingly and include:
ClinicalData _YaleTrainingCV_ShortTermFollowUP.csv
ClinicalData _YaleTrainingCV_LongTermFollowUP.csv
ClinicalData _YaleIndependent_ShortTermFollowUP.csv
ClinicalData _YaleIndependent_LongTermFollowUP.csv
ClinicalData _Geisinger_LongTermFollowUP.csv

2. Experimental Design, Materials and Methods

2.1. Patient Population

The dataset consists of patients from two institutions: Yale New Haven Health (New Haven, CT, USA; $n = 597$) and Geisinger Health (Danville, PA, USA; $n = 232$). Yale subjects were identified from the Yale stroke center registry between 1/1/2014 and 10/31/2020, and Geisinger subjects were identified from the Geisinger stroke center registry between 1/1/2016 and 12/31/2019. As depicted in the related research article Supplementary Fig. 1, subjects were included if they (1) suffered an anterior circulation large vessel occlusion (LVO) stroke – including internal carotid artery (ICA) or middle cerebral artery (MCA) M1 or M2 segments, (2) had CTA source images with slice thickness $\leq$1 mm, (3) underwent endovascular thrombectomy (ET), and (4) had modified Rankin Scale (mRS) assessment of functional outcome recorded at discharge or at 3-month follow-up. Patients were excluded if they had (1) any simultaneous posterior circulation LVO, (2) poor quality CTA not amenable to analysis (due to motion, metal artifact, or scanner-based artifacts), or (3) missing admission clinical information.

2.2. Image Processing and Radiomics Feature Extraction

We modified the brain extract tool (BET) from FSL software (http://www.fmrib.ox.ac.uk/) to perform skull-stripping of each patient's admission CTA [4]. Next, we applied FLIRT from the FSL toolbox to co-register each CTA to the Montreal Neurological Institute (MNI)−152 brain space. We used the brain stroke atlas to generate bilateral MCA territory masks in MNI-152 space [5]. Then, bilateral MCA territory masks were reverse registered to the native CTAs.

Trilinear interpolation was used to resample all CTA images within MCA territory masks to an isotropic $1 \times 1 \times 1$ mm voxel spacing. This ensured rotational invariance of texture features [6–8]. All images were normalized by centering voxel values at the mean with standard deviation from the image. To ensure exclusion of calcified plaques or remaining skull tissue, only voxels within a 1–500 Hounsfield unit (HU) range were included in analysis. To compensate for differences in intravenous bolus timing among different CTA scans, the voxel values in each patient was normalized to the mean attenuation of the scan during radiomics feature extraction process. We applied high- and low-pass filters in each spatial direction ("coif-1″ wavelet transform [3]) and the "edge-enhancement" Laplacian of Gaussian (LoG) filter (with "sigma" settings of 2,4,and 6 mm [3]). We then extracted one set of 1116 "first-order" and "texture-matrix" radiomics features per patient from the single volume of interest (VOI), combining right and left MCA territories [3]. We utilized a custom Pyradiomics version 2.1.2 pipeline [3] to complete the steps of preprocessing, derivative image generation, and feature extraction. Supplementary Table 1 of the related research article [1] describes the first-order and texture-based features.

2.3. Machine Learning Framework

Six dimensionality reduction strategies and six machine learning classifiers appropriate for application to radiomics data are listed in the Analyzed Data file Table 1 and described in detail in the related research article supplement [1], along with their programming packages. Each combination of these dimensionality reduction strategies and machine learning classifiers were used to create 36 candidate models for prediction of LVO stroke patient outcome in the related research article [1].

The dimensionality reduction methods include: hierarchical clustering, maximum relevance minimum redundancy filtering, no feature selection, principal component analysis, Pearson correlation-based redundancy reduction with mutual information maximization filter, and RIDGE regularized logistic regression for feature selection. The machine learning classifiers include: elastic net regularized logistic regression, Naïve Bayes, random forest, support vector machine with radial kernel, support vector machine with sigmoid kernel, and extreme gradient boosting. The hyperparameters, their ranges, and tuning repetition counts used for each machine learning classifier are described in the related research article Supplementary Table 2 [1]. Detailed explanation of the machine learning training and validation methodologies can be found in the methods section of the supplementary research article [1].

Ethics Statements

Institutional Review Board approval was obtained for data collection (Yale University protocol number 2000024296), with informed consent waived at respective institutes due to the retrospective nature of our study. All procedures followed were in accordance with institutional guidelines and the Declaration of Helsinki.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Dr. Christopher G. Filippi receives consulting honoraria from Syntactx, Inc; minority stockholder in Avicenna.ai; and receives research funding from the National Multiple Sclerosis Society.

Dr. Christoph Griessenauer receives research funding from Medtronic and Penumbra and consulting honoraria from Stryker and MicroVention.

Dr. Philipp Hendrix receives salary support from Medtronic, which was used to support this work.

Dr. Kevin Sheth receives grant support from Novartis, Biogen, Bard, Hyperfine and Astrocyte. He also reports equity interests in Alva Health.

Data Availability

Dataset on acute stroke risk stratification from CT angiographic radiomics (Original data) (GitHub).

CRediT Author Statement

Emily W. Avery: Conceptualization, Investigation, Data curation, Methodology, Writing – original draft, Writing – review & editing, Visualization; **Jonas Behland:** Investigation, Data curation; **Adrian Mak:** Data curation, Investigation, Writing – review & editing; **Stefan P. Haider:** Methodology, Software, Writing – review & editing; **Tal Zeevi:** Methodology, Software, Writing – review

& editing; **Pina C. Sanelli:** Supervision, Writing – review & editing; **Christopher G. Filippi:** Investigation, Data curation, Writing – review & editing; **Ajay Malhotra:** Conceptualization, Supervision, Writing – review & editing; **Christoph J. Griessenauer:** Investigation, Data curation, Writing – review & editing; **Ramin Zand:** Investigation, Data curation, Writing – review & editing; **Philipp Hendrix:** Investigation, Data curation, Writing – review & editing; **Vida Abedi:** Investigation, Data curation, Writing – review & editing; **Guido J. Falcone:** Investigation, Writing – review & editing; **Nils Petersen:** Investigation, Writing – review & editing; **Lauren H. Sansing:** Investigation, Writing – review & editing; **Kevin N. Sheth:** Investigation, Writing – review & editing; **Seyedmehdi Payabvash:** Conceptualization, Supervision, Data curation, Writing – original draft, Writing – review & editing.

Acknowledgments

Funding: This work was supported by the National Institutes of Health [grant numbers K76AG059992, R03NS112859, P30AG021342, KL2 TR001862, U24NS107215, U24NS107136, U01NS106513, R01NR018335, R01NS095993, R01NS097728, K23NS118056]; the American Heart Association [grant numbers 18IDDG34280056, 17MCPRP33460188, 17CSA33550004]; the Doris Duke Charitable Foundation [grant number 2020097]; the Radiological Society of North America [grant number A129581]; the American Society of Neuroradiology; and Yale University.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2022.108542.

References

[1] E.W. Avery, J. Behland, A. Mak, S.P. Haider, T. Zeevi, P.C. Sanelli, C.G. Filippi, A. Malhotra, C.C. Matouk, C.J. Griessenauer, R. Zand, P. Hendrix, V. Abedi, G.J. Falcone, N. Petersen, L.H. Sansing, K.N. Sheth, S. Payabvash, CT angiographic radiomics signature for risk stratification in anterior large vessel occlusion stroke, Neuroimage Clin. 34 (2022) 103034, doi:10.1016/j.nicl.2022.103034.

[2] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillion-Robin, S. Pieper, H. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (21) (2017) e104–e107, doi:10.1158/0008-5472.CAN-17-0339.

[3] J.J.M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillon-Robin, S. Pieper, H.J.W. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (7) (2017) e104–e107, doi:10.1158/0008-5472.CAN-17-0339.

[4] J. Muschelli, N.L. Ullman, W.A. Mould, P. Vespa, D.F. Hanley, C.M. Crainiceanu, Validated automatic brain extraction of head CT images, Neuroimage 114 (2015) 379–385, doi:10.1016/j.neuroimage.2015.03.074.

[5] Y. Wang, J.M. Juliano, S.L. Liew, A.M. McKinney, S. Payabvash, Stroke atlas of the brain: voxel-wise density-based clustering of infarct lesions topographic distribution, Neuroimage Clin. 24 (2019) 101981, doi:10.1016/j.nicl.2019.101981.

[6] S.P. Haider, A. Mahajan, T. Zeevi, P. Baumeister, C. Reichel, K. Sharaf, R. Forghani, A.S. Kucukkaya, B.H. Kann, B.L. Judson, M.L. Prasad, B. Burtness, S. Payabvash, PET/CT radiomics signature of human papilloma virus association in oropharyngeal squamous cell carcinoma, Eur. J. Nucl. Med. Mol. Imaging 47 (13) (2020) 2978–2991, doi:10.1007/s00259-020-04839-2.

[7] S.P. Haider, T. Zeevi, P. Baumeister, C. Reichel, K. Sharaf, R. Forghani, B.H. Kann, B.L. Judson, M.L. Prasad, B. Burtness, A. Mahajan, S. Payabvash, Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma, Cancers (Basel) 12 (7) (2020), doi:10.3390/cancers12071778.

[8] S.P. Haider, K. Sharaf, T. Zeevi, P. Baumeister, C. Reichel, R. Forghani, B.H. Kann, A. Petukhova, B.L. Judson, M.L. Prasad, C. Liu, B. Burtness, A. Mahajan, S. Payabvash, Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics, Transl. Oncol. 14 (1) (2020) 100906, doi:10.1016/j.tranon.2020.100906.