

Article

# Defining the Age of Young Ischemic Stroke Using Data-Driven Approaches

Vida Abedi <sup>1,2,†</sup> , Clare Lambert <sup>3,†</sup> , Durgesh Chaudhary <sup>4,5,†</sup> , Emily Rieder <sup>6</sup> , Venkatesh Avula <sup>1</sup> ,  
Wenke Hwang <sup>2</sup> , Jiang Li <sup>1</sup>  and Ramin Zand <sup>4,5,\*</sup> 

<sup>1</sup> Department of Molecular and Functional Genomics, Weis Center for Research, Geisinger Health System, Danville, PA 17822, USA

<sup>2</sup> Department of Public Health Sciences, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA

<sup>3</sup> Department of Neurology, Yale New Haven Hospital, New Haven, CT 06510, USA

<sup>4</sup> Geisinger Neuroscience Institute, Geisinger Health System, Danville, PA 17822, USA

<sup>5</sup> Department of Neurology, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA

<sup>6</sup> Geisinger Commonwealth, School of Medicine, Scranton, PA 18509, USA

\* Correspondence: [ramin.zand@gmail.com](mailto:ramin.zand@gmail.com); Tel.: +1-(717)-531-1804; Fax: +1-(717)-531-0384

† These authors contributed equally to this work.

**Abstract: Introduction:** The cut-point for defining the age of young ischemic stroke (IS) is clinically and epidemiologically important, yet it is arbitrary and differs across studies. In this study, we leveraged electronic health records (EHRs) and data science techniques to estimate an optimal cut-point for defining the age of young IS. **Methods:** Patient-level EHRs were extracted from 13 hospitals in Pennsylvania, and used in two parallel approaches. The first approach included ICD9/10, from IS patients to group comorbidities, and computed similarity scores between every patient pair. We determined the optimal age of young IS by analyzing the trend of patient similarity with respect to their clinical profile for different ages of index IS. The second approach used the IS cohort and control (without IS), and built three sets of machine-learning models—generalized linear regression (GLM), random forest (RF), and XGBoost (XGB)—to classify patients for seventeen age groups. After extracting feature importance from the models, we determined the optimal age of young IS by analyzing the pattern of comorbidity with respect to the age of index IS. Both approaches were completed separately for male and female patients. **Results:** The stroke cohort contained 7555 ISs, and the control included 31,067 patients. In the first approach, the optimal age of young stroke was 53.7 and 51.0 years in female and male patients, respectively. In the second approach, we created 102 models, based on three algorithms, 17 age brackets, and two sexes. The optimal age was 53 (GLM), 52 (RF), and 54 (XGB) for female, and 52 (GLM and RF) and 53 (RF) for male patients. Different age and sex groups exhibited different comorbidity patterns. **Discussion:** Using a data-driven approach, we determined the age of young stroke to be 54 years for women and 52 years for men in our mainly rural population, in central Pennsylvania. Future validation studies should include more diverse populations.

**Keywords:** ischemic stroke; data science; machine-learning; electronic health records; EHR



**Citation:** Abedi, V.; Lambert, C.; Chaudhary, D.; Rieder, E.; Avula, V.; Hwang, W.; Li, J.; Zand, R. Defining the Age of Young Ischemic Stroke Using Data-Driven Approaches. *J. Clin. Med.* **2023**, *12*, 2600. <https://doi.org/10.3390/jcm12072600>

Academic Editor: Peter Sporns

Received: 22 February 2023

Revised: 15 March 2023

Accepted: 21 March 2023

Published: 30 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

From 1989 to 2008, there has been a 36–38% increase in “young strokes” in the United States [1]. However, the definition of “young stroke” differs from source to source, making it difficult to ascertain the incidence of ischemic stroke (IS) in young adults. Up to 10% of first-time ISs occur in people under the age of 45 [2], and nearly one-quarter of strokes occur in the working-age population [3]. Despite a fall in the overall prevalence of IS over the last decade, there has been a rise in “young strokes” [2]. More rigorous management of traditional IS risk factors, such as diabetes, hypertension, hypercholesterolemia, and

smoking, may have contributed to a decline in whole-population stroke rates in older populations [4]; meanwhile, the prevalence of rarer etiologies that plague predominantly younger cohorts, such as hypercoagulable states, vasculitis, and genetic causes, Ref. [5] remain the same. Additionally, the prevalence of traditional risk factors may be shifting toward younger cohorts over time [6].

There is no standard definition of “young stroke”, and most studies have used an arbitrary upper cutoff between 45 and 59 years [7–13]. The age cutoff is clinically meaningful, since the uncommon stroke etiologies are more frequent among the younger population, and they may need a more comprehensive workup. The age cutoff is also epidemiologically meaningful to monitor the IS risk factors and incidence among different age brackets, and investigate rare causes, especially genetic predispositions. In our previous study on young stroke patients’ outcomes, we used multiple upper-age cut-points (43.7, 49.0, and 55.0 years) to define the young IS cohort [14]. The current study aimed to determine the age cut-point for defining “young stroke” in men and women, using a data-driven approach and comprehensive clinical data from a large, mainly rural, population in the United States. We further investigated the rare and traditional stroke risk factors among the young IS cohorts.

## 2. Methods

The study was conducted and reported according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines [15]. TRIPOD provides a set of recommendations for the reporting of studies developing, validating, or updating a prediction model to improve transparency and reproducibility. The Supplemental Material (Table S1) includes the TRIPOD checklist. The study was reviewed and approved by the Geisinger Institutional Review Board to meet “non-human subject research” for using de-identified information.

We used two parallel approaches to determine and corroborate the age cut-point for young stroke. In the first approach, we used diagnosis codes based on ICD-9/10-CM (based on the entire comorbidity profile) from IS patients to group patient comorbidities, computed similarity scores between every patient pair, and determined the optimal age cut-point for defining young stroke by analyzing the trend of patient similarity with respect to their respective comorbidities for different age at IS. A similarity score was used to measure how every two patients were comparable based on their risk profiles. For the second approach, we used the IS patient cohort and controls (patients without IS) to build three sets of machine-learning (ML) models to classify patients with IS for 17 different age groups. We then extracted the feature importance from the models and determined the optimal age cut-point of young stroke by analyzing the pattern of comorbidity for different groups of patients based on the IS age of onset. The second strategy used clinical data based on common and rare vascular risk factors. Both approaches were performed separately for male and female patients.

### 2.1. Data Source and Patient Population

Consecutive patient-level structured data from 13 hospitals of a large health system (Geisinger) in Pennsylvania, United States, from September 2003 to May 2019, were used for this study. Patients were included in the study if they were at least 18 years old at stroke onset and had no prior history of hemorrhagic or ischemic stroke. The patients were labeled as IS if they had an encounter at the Geisinger emergency department (ED) with a discharge diagnosis of ischemic stroke (IS), a brain MRI during their encounter, and an encounter duration of 24 h or more. The patients were labeled as control (without IS) if they had an ED encounter at Geisinger and an inpatient stay of at least 24 h with brain imaging (CT or MRI), but did not have a discharge diagnosis of stroke, transient ischemic attack (TIA), or intracerebral hemorrhage (ICH). For the second approach, different groups were created based on the patient’s age at the index encounter. All meta-data and code developed

for this study are available in a GitHub repository: <https://github.com/TheDecodeLab/YoungStroke> (accessed on 20 March 2023).

## 2.2. Data Variables

For the first approach, all the diagnosis codes, based on ICD-9-CM/ICD-10-CM, prior to index IS were extracted for all the IS patients; these codes were mapped to 534 Phecodes [16]. This data-driven strategy reduced the need to select associated factors manually. For the second approach, literature and experts were consulted, and a total of 23 variables (Table S2) were selected; this list was constructed according to the International Pediatric Stroke Study Classifications [17], and other sources [18–22].

The variables included the patient's past medical history, demographics, and clinical data; all the variables were available and extracted (see Supplemental Material Table S3 for relevant ICD-9-CM/ICD-10-CM codes) from Geisinger's electronic health record (EHR) data warehouse. Some variables were obtained only at the time of index stroke, others had a three-month buffer window after index stroke, and some were obtained if identified at any time in the patient's medical record. The data variables and the extraction time frame are provided in Table S2. Variables that were given a three-month buffer window were those that could take time after hospital discharge to reach a definitive diagnosis, such as "hypercoagulable state". Variables present at any time included those that afflicted patients lifelong, such as fibromuscular dysplasia.

## 2.3. Data Pre-Processing and Imputation

Body Mass Index (BMI) information closest to the index ED encounter was dichotomized and used as a binary variable (normal: if BMI  $\leq$  25 or overweight: if BMI > 25). The BMI sustained 2.3% missingness, and an imputation was performed using the MICE (Multivariate Imputation by Chained Equations) package [23]. We have previously shown that, for missingness in the range in this study, MICE is an adequate choice for EHR [24,25].

## 2.4. Statistical Analysis

All binary variables were presented as count and percentage. The prevalence of each variable within the cases was compared to controls, using a Pearson's chi-squared test or Fisher's exact test. A *p*-value of less than 0.05 was considered significant for all analyses. All statistical analyses were performed using R version 4.0.3 [26].

## 2.5. First Approach: Phenotype Modeling

As mentioned above, the ICD9-CM/ICD10-CM codes of patients with first-time IS were mapped to Phecode using Phecode Map 1.2 (R package 'PheWAS') [16]. After grouping patients by sex, a similarity score between every pair of patients was calculated using the matrix of Phecodes as the input, and cosine similarity as the scoring metric. The median similarity score for each patient was calculated and plotted versus the age of IS onset. The scatter plot was then analyzed to identify a trend of a similar pattern, using locally weighted regression (Loess regression) to fit a smoothed line. The estimated cut-point for approximating the age of young stroke was determined as the age best separating the patient similarity score by standardized Wilcoxon statistic, using the R package 'maxstat'.

## 2.6. Second Approach: Machine Learning

We used three algorithms—logistic regression (generalized linear models, GLM), in [27] extreme gradient boosting (XGB), [28] and random forest (RF) [29]—to train classifiers based on IS (cases) and non-IS patients (controls) for different age brackets. Following the classification, we extracted the feature importance, based on the 23 variables, for the different age groups; the models were created separately for males and females. The relative feature importance of each variable in each age bracket, and for males and females, was used to identify the changing comorbidity patterns. We created 102 models, based on 3 algorithms, 17 age brackets, and 2 sexes. The 17 age brackets included IS index age below

40, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 65, 70, 75, and 80. Given that most studies used an age bracket in the range of 45–55, we designed our study to have a higher resolution for this age range.

We split the data for each group into training (80%) and testing sets (20%). A parameter grid was built to train the model with five-fold repeated cross-validation, with five repeats. Model-tuning was performed by an automatic grid search, with five different values to try for each algorithm parameter randomly. Finally, 20% testing set data were used to calculate the model AUROC as a metric of model performance. The average feature importance was extracted from the fine-tuned models in each age and sex group. A heatmap was also generated to help visualize the patterns of change over time for the 23 variables.

A comparison of comorbidity patterns was performed by calculating the similarity measure between feature importance vectors for different age and sex groups. The similarity of the pattern of comorbidities among different groups can be a measure of how different these cohorts are, as the age of IS onset is varied from 40 to 80 years. The similarity was calculated using the cosine similarity measure. Finally, the non-parametric Wilcoxon statistic test was performed on the cosine similarity scores for different age brackets to determine the age cut-point, using the R package ‘*maxstat*’.

### 3. Results

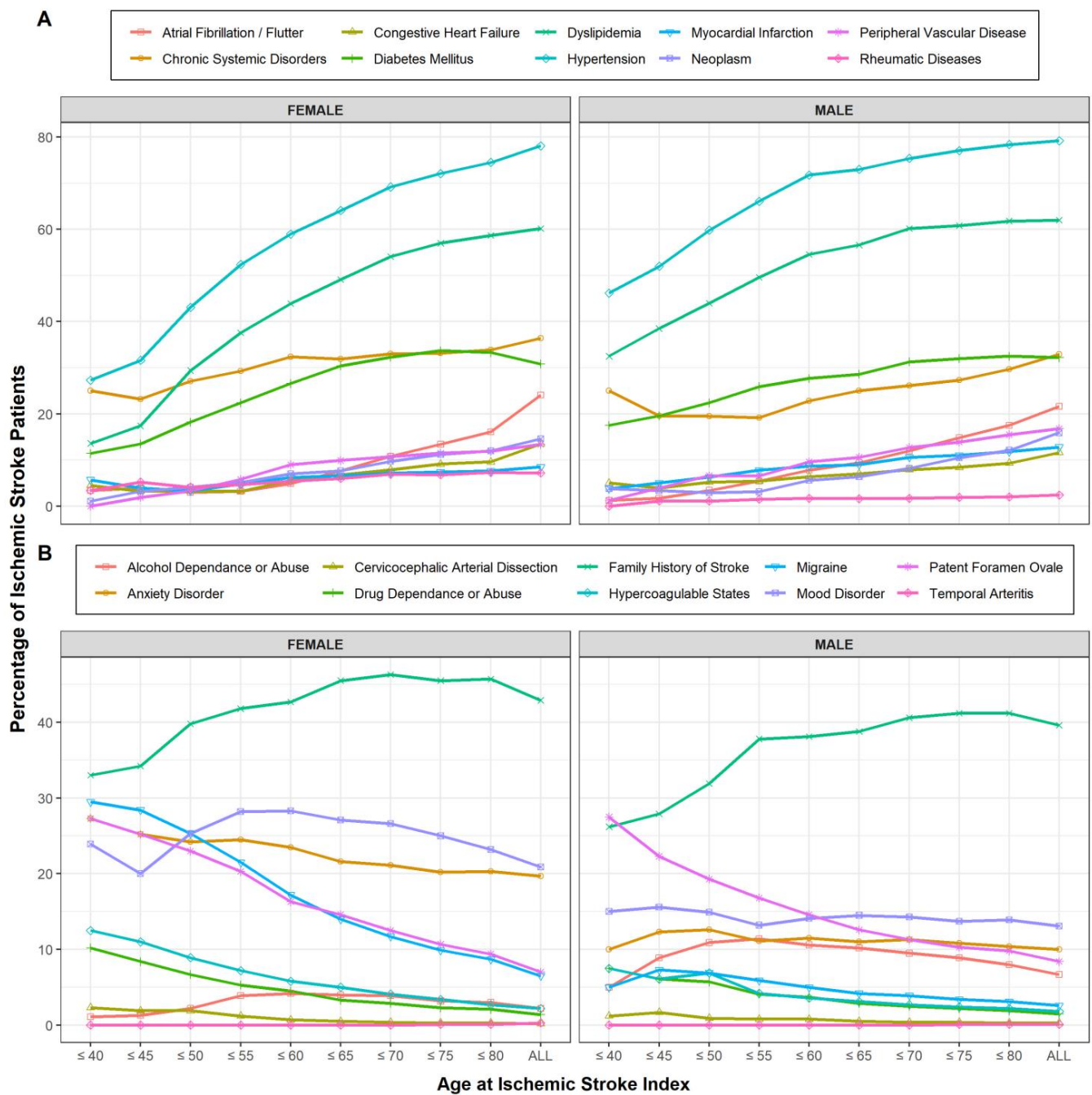
#### 3.1. Patient Population and Characteristics

The stroke cohort comprised 7555 IS patients who presented to Geisinger from 2003 to 2019, and met our inclusion criteria. The control cohort included 31,067 adult patients who presented to the Geisinger ED during the same period, and met the inclusion criteria for our control group. Females and males were analyzed separately. A total of 3892 males and 3663 females had IS during the study period; 66.3% of males and 50.1% of females were under the age of 75 at the first-time stroke, whereas 15.8% of males and 11.8% of females were under the age of 55 (Tables S4 and S5).

The prevalence of certain risk factors did not differ significantly between cases (IS) compared to controls (without IS). For example, in the female cohorts, alcohol abuse or dependence, neoplasm, rheumatic disease, mood disorder, and giant cell arteritis (GCA) demonstrated no statistically significant difference in prevalence between cases and controls for any age brackets. In female patients, drug dependence or abuse was significantly different in the age brackets of <40 and <45, whereas current smoking was only different in the <80 and the all-ages brackets. For males, current smoking and GCA were the only variables that showed no statistical difference between cases and controls.

Generally, three main risk factors patterns were measured: (1) an increase in prevalence with increasing age, (2) a decrease in prevalence with increasing age, and (3) relative steadiness with increasing age. For example, in women, migraine (29.5% of <40 versus 9.7% of <80) and PFO (27.3% of <40 versus 9.4% of <80) demonstrated a declining prevalence. In men, PFO followed the same pattern (27.5% in <40 versus 9.8% in <80). In women, hypertension (7.3% in <40 versus 74.5% in <80), dyslipidemia (13.6% in <40 versus 58.7% in <80), and diabetes (11.4% in <40 versus 33.3% in <80) are examples of variables that demonstrated an opposite pattern, whereby the prevalence increased with age at IS onset. In men, hypertension (46.2% in <40 versus 78.9% in <80), dyslipidemia (32.5% in <40 versus 61.8% in <80), neoplasm (3.8% in <40 versus 12.1% in <80), and MI (3.8% in <40 versus 11.8% in <80) followed this pattern. Some variables, such as BMI, arteriopathies, cervicocephalic dissection in women, mood disorders, smoking, and cervicocephalic dissection in men, did not fluctuate with respect to IS age of onset. A list of variables and the respective percentages for the different age brackets in both males and females are summarized in Table S4 (females) and Table S5 (males). Figure 1A,B shows the changing prevalence of comorbidities with respect to the age of IS onset.

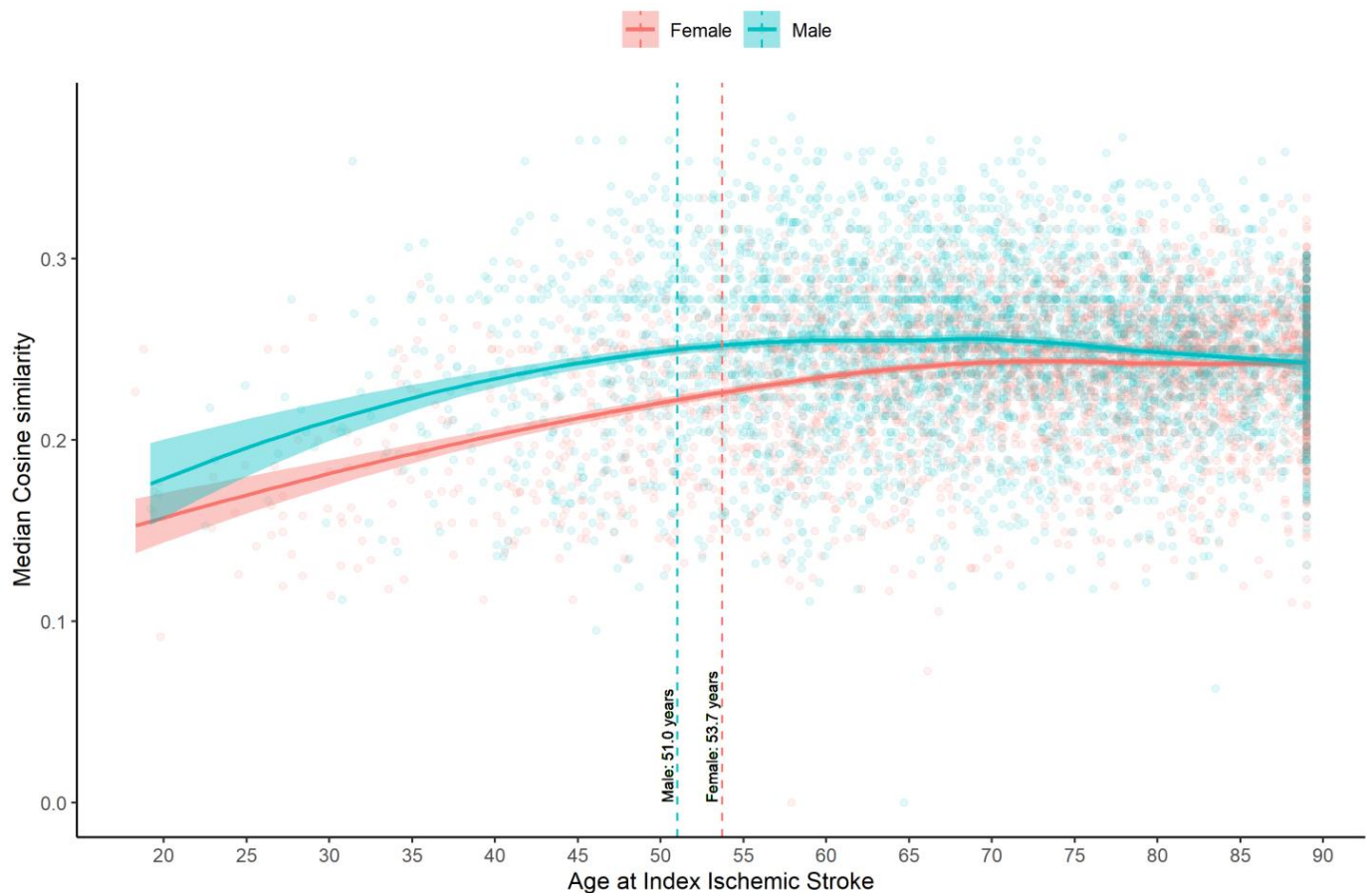




**Figure 1.** Line graph showing the prevalence of comorbidities as a percentage of the group total (*y*-axis) with respect to the age bracket increments (*x*-axis). (A) Line graph of first half of risk factors, (B) Lines graph of the second half of risk factors. All risk factors are spread over two graphs for clarity given high number of risk factors.

### 3.2. Estimating the Age Cut-Point for Young IS Using Phenotype Modeling (Approach 1)

The ICD-9-CM/ICD-10-CM codes of IS patients up to the index stroke date were mapped to 534 Phecodes. Apart from cerebrovascular disease, the top five Phecodes groups in the IS patients were hypertension (74%), disorders of lipid metabolism (61.9%), neurological disorders (38.1%), diabetes mellitus (31.6%), and cardiac dysrhythmias (30.4%). After calculating the cosine similarity based on the Phecodes variable, the age cut-point was estimated to be 53.7 years in female IS patients and 51.0 years in male IS patients (Figure 2).



**Figure 2.** Scatter plot of median cosine similarity and age at index ischemic stroke date with a smoothed line, using loess regression.

### 3.3. Estimating the Age Cut-Point for Young IS Using Machine Learning (Approach 2)

Models can be developed to predict stroke using clinical data. In total, three algorithms (LR, XGB, and RF) were built for each age group (<40, <45, <46, <47, <48, <49, <50, <51, <52, <53, <54, <55, <60, <65, <70, <75, <80, and all), and for males and females separately, resulting in 102 models (three algorithms  $\times$  17 age brackets  $\times$  two sexes). Supplemental Figure S1 summarizes the model AUROC and accuracy of each age/sex group in classifying stroke patients. Model parameters are summarized in Supplemental Material Table S6. In general, as the number of patients increased with the index age, the model AUROC increased steadily, reaching above 0.85. Model accuracy was less affected; however, the overall trend was slightly higher in models developed for females (overall accuracy above 0.8 was observed for all age brackets in the female group). Feature importance, extracted from these models for each age bracket and gender, was used to study the pattern of comorbidities, as the patients' age at the index increased.

Feature importance, extracted from ML-based models, can be used to estimate the age of "young stroke." The pattern of feature importance can be visualized in a heatmap (Figure 3). To quantify the patterns observed in Figure 3, the similarity between the contribution of risk factors to the classification was measured for each age bracket using the cosine measure. More specifically, comparisons were made between the less-than age groups and the entire IS population. For instance, feature importance vectors for females <40 years were compared with females in all age groups. The resulting trends (trend of similarity measures for males and females), presented in Figure 4, showed the pattern of change in the associated factors based on the ML-based models. Our ML-based models showed that in females, the three models led to an age of 53 (based on GLM), 52 (based on RF), and 54 (based on XGB) for defining the age cutoff of young IS. In comparison, this

pattern changed slightly for male patients: 52 years based on GLM and RF, and 53 based on the XGB model. These results corroborated our data-driven strategy based on phenotype modeling (approach 1) in defining the age of young stroke.

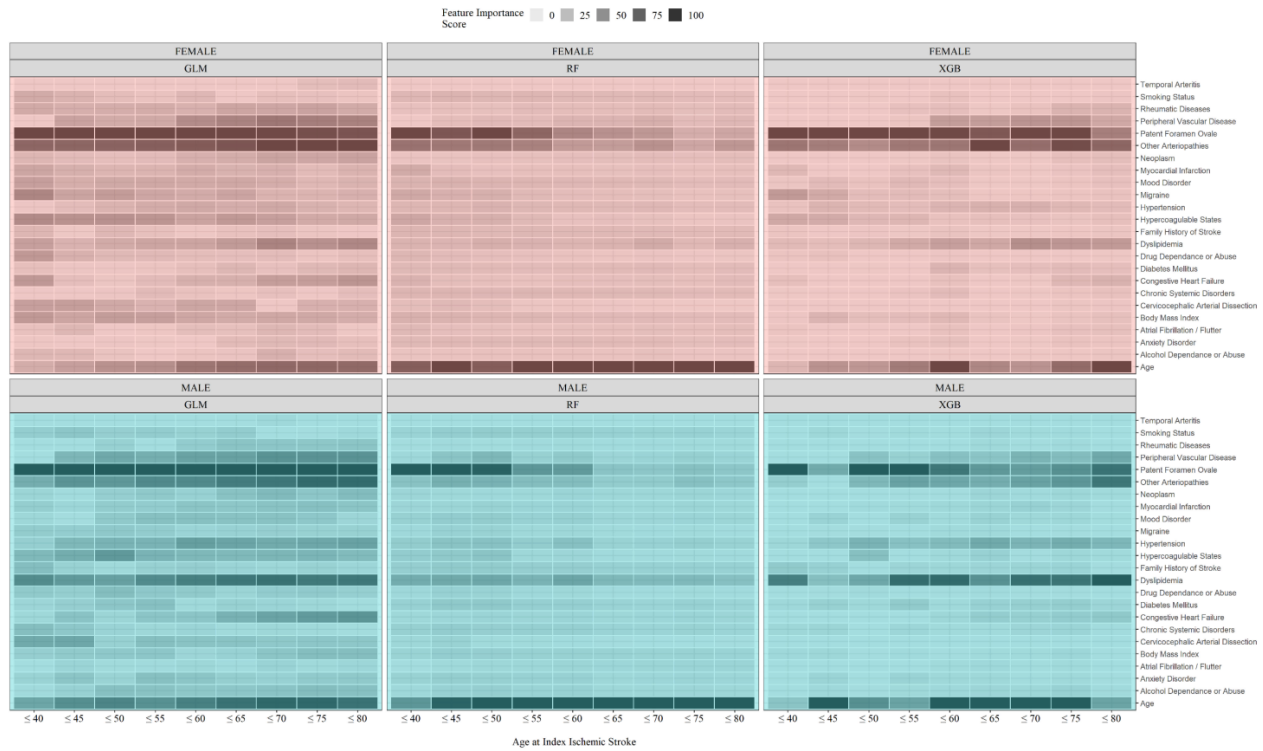


Figure 3. Average feature importance of the variables used by GLM, RF, and XGB models.

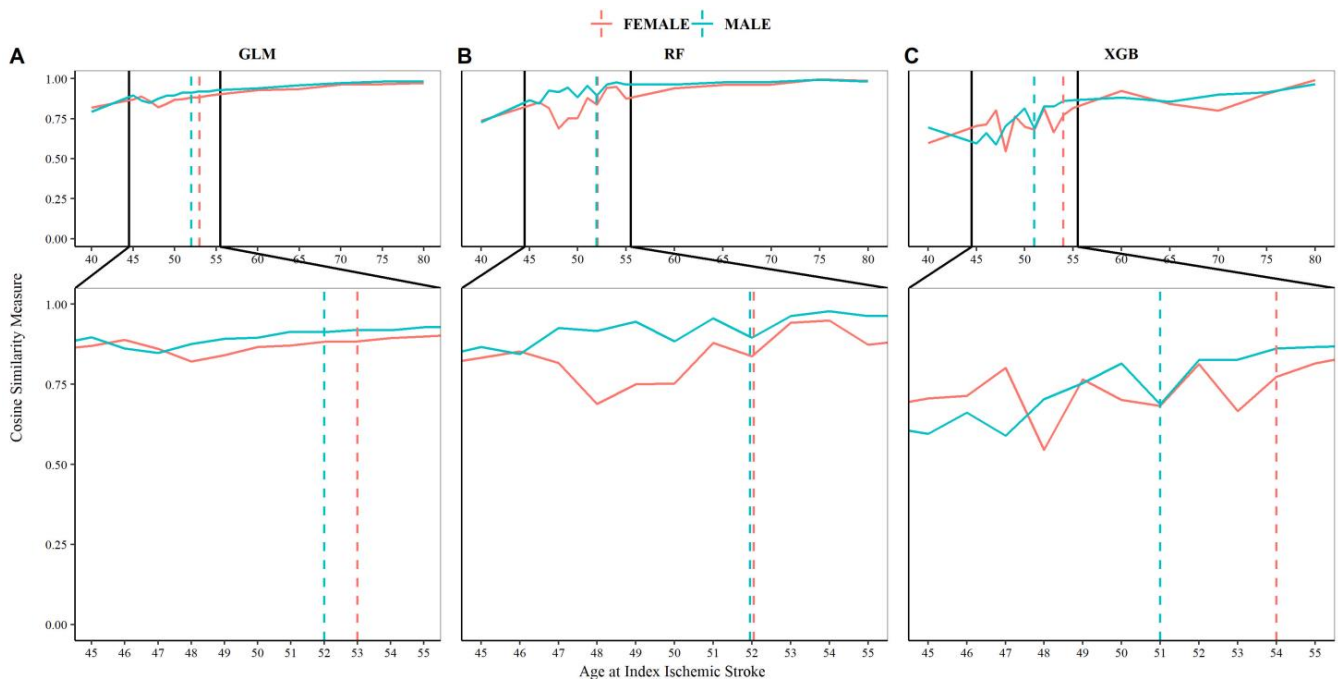


Figure 4. The trend of cosine similarity versus age at index ischemic stroke. Lower cosine similarity scores were indicative of increased dissimilarity, suggesting a change in risk factor profile across that age cutoff. (A) Cosine similarity trend for GLM model. (B) Cosine similarity trend for RF model. (C) Cosine similarity trend for XGB model.

## 4. Discussion

### 4.1. “Young Stroke” Cutoff

This is the first study using a data-driven strategy to identify the most appropriate age for young stroke in male and female patients, using a large population-based cohort. Two different approaches, based on phenotype analysis and supervised ML modeling, corroborated that young stroke has an optimal age cut-point, which is also slightly different in male and female patients. A more precise cut-point is 54 years for women and 52 years for men. This finding is important when analyzing clinical data or determining young stroke for epidemiological or genetic studies. Past literature has not defined an age cutoff that differentiates “young stroke” from the traditional stroke that occurs most often in elderly people, Ref. [30] and arbitrary age cutoffs have been applied across various studies [31]. A summary of important studies looking at young stroke, highlighting the variability of chosen age cutoffs, is included in Table S7.

### 4.2. Vascular Risk Factors

In our cohort, young patients had lower rates of traditional and higher rates of rare risk factors compared to older patients. For example, hypertension and dyslipidemia existed at a lower prevalence in younger age brackets than in older; however, all age brackets demonstrated higher rates than controls. This observation was consistent when analyzing the feature importance extracted from the ML models. Kissella et al. (2012) highlighted an increase in self-reported vascular risk factors among young people [32]. A recent European study demonstrated that smoking, dyslipidemia, and hypertension were the most common risk factors in people under 49 years of age with IS, versus coronary artery disease, atrial fibrillation, and diabetes, which were more characteristic of those over the age of 61 [33]. Overall, the reason why young people are suffering an increasing number of strokes remains unclear. An Australian study identified “cryptogenic” as the listed etiology in nearly half of the strokes occurring in patients aged 18–50 years, compared to only 18.1% in those over the age of 61 [34]. In our cohort, some other well-known vascular risk factors, such as obesity and family history of stroke or MI, remained high across all age brackets. It is conceivable that younger patients with vascular risk factors may be exponentially more at risk of having a stroke if their presentation is compounded by rare etiologies that are relatively less frequent in older individuals, such as migraine, hypercoagulable states, and PFO; however, the reason for the increasing prevalence of young stroke is yet to be elucidated.

### 4.3. Rarer Stroke Etiologies and Non-Traditional Risk Factors

Interestingly, some known risk factors for stroke did not differ significantly between cases and controls in both males and females across all or most age brackets. These risk factors included elevated BMI, current smoking, cancer, rheumatic disease, alcohol abuse/dependence, drug abuse/dependence, and GCA. Some of the lack of difference observed in these risk factors may be related to the way variables were captured; for example, cocaine use is a known risk factor for stroke secondary to vasospasm, Ref. [35] but this may have been diluted when abuses of all drugs were condensed into a single variable. The fact that rates of smoking did not differ significantly between cases and controls is not clear; however, that could be related to better reporting of smoking status in EHR and an overall decrease in rates of smoking [36,37].

Anxiety and mood disorders (depression and bipolar disorder) were included in our analysis, given the increasing data supporting the association between mental illness and an increased risk of stroke [21,22]. For women, there was no difference in mood disorder prevalence between cases and controls; however, for men, both pre-index IS mood disorders and anxiety disorder had a significantly higher prevalence in those with stroke compared to control. This is also observed when analyzing the feature importance extracted from the ML models (Figure 3). While this finding does not imply causation, it highlights that psychological distress may play more of a role in stroke than previously



thought. While there is good evidence surrounding post-stroke depression, Ref. [38] there is less information on how having a mood disorder may pre-dispose patients to stroke, particularly if the respective psychiatric diagnosis affects patients' ability to seek out and engage with health-promotion efforts, which in turn highlights the importance of personalized and targeted management in stroke-prevention efforts.

#### 4.4. Gender, Genetic, and Environment Differences

Several other non-age-related factors may play into the prevalence of "young strokes", including environment and sex. Some studies have suggested that individuals residing in rural areas have worse stroke outcomes than in urban areas, and are plagued by more stroke risk factors [39]. There is a relative lack of data on young people with stroke in rural settings, making the present investigation based on a rural region in the United States a unique source of information. These patients may be unintentionally excluded from primary prevention efforts due to a lack of access to services or the misconception that their risk of stroke is limited due to their age [40,41].

Female "young stroke" is unique, given fluctuant hormone levels present in women during the peri-menopausal stage [42]. A European study identified dyslipidemia (51.6% in males versus 37.9% in females,  $p < 0.001$ ), smoking (54.3% in males versus 41.4% in females,  $p < 0.001$ ), and coronary heart disease (7.8% in males versus 3.8% in females,  $p < 0.001$ ) as examples of risk factors that differed significantly between genders in patients with stroke, who were under the age of 49 [33]. There was no absolute sex difference in this cohort in terms of family history of stroke, hypertension, past transient ischemic attack (TIA), diabetes, peripheral arterial disease (PAD), AF, and heart failure (HF) [33]. However, many women will not have undergone menopause at age 49. A different pattern of gender difference might have merged in the 50–60 years old category, had they been included in the study [42,43]. Lisabeth et al. (2010) identified early menopause as a significant risk factor for stroke in women after the age of 60; however, limited data exist to suggest how the peri-menopausal risk factor profile differs from pre-menopausal women, and men [44].

There is a growing body of research looking at a genetic predisposition to stroke. Genetic causes of stroke may disproportionately affect young people. Interestingly a recent large-scale meta-analysis genome-wide association study (GWAS) identified the A1 blood haplotype variation at the ABO locus to be associated with early-onset IS compared to later-onset strokes [7]. The mechanism of this is hypothesized to be related to thrombosis rather than atherosclerosis. In that GWAS study, the age cut-point to define young stroke was 59 years for male and female patients.

#### 4.5. Study Limitations, Strengths, and Future Directions

This study has several limitations, and the results must be interpreted within these constraints. The EHR data used in model development were rich and comprehensive; however, there is an inherent noise associated with EHR, including selection bias. The patient population is based on the regional demographics in central and northeast Pennsylvania, a predominantly Caucasian patient population. This study also had several strengths: using a large cohort from an integrated healthcare system and a data-driven strategy coupled with supervised modeling to corroborate the findings.

When considering sex as a biological variable, the age for defining young stroke is also important. We have shown that a more precise definition can be reached if the analysis is done in parallel for male and female patients. Improved patient stratification for future investigations is critical in improving care and management for patients suffering a stroke at a younger age. Given the low proportion of strokes experienced by younger patients at any given teaching hospital, a meta-analysis may be the only way to truly elucidate what factors are driving the increasing stroke prevalence in this age group, and that is why having a consensus on defining "young stroke" is an essential first step.

## 5. Conclusions

This is the first study to leverage large data and an innovative data-driven strategy to identify the most appropriate age to identify young male and female IS patients. Our data suggest that younger patients presented higher rates of traditional, as well as rare, risk factors than controls. In comparison, older patients displayed higher rates of traditional risk factors. A more precise cut-point in defining young stroke is 54 years for women and 52 years for men in our, mainly rural, population in central Pennsylvania. Future validation studies should include more diverse populations.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jcm12072600/s1>.

**Author Contributions:** Conceptualization, V.A. (Vida Abedi) and R.Z.; Methodology, V.A. (Vida Abedi), C.L., D.C., E.R., V.A. (Venkatesh Avula) and J.L.; Validation, V.A. (Vida Abedi), C.L., D.C.; Formal analysis, V.A. (Vida Abedi), C.L. and D.C.; Investigation, V.A. (Vida Abedi), C.L., D.C., V.A. (Venkatesh Avula) and R.Z.; Resources, V.A. (Vida Abedi) and R.Z.; Data curation, V.A. (Vida Abedi), C.L., D.C., E.R., V.A. (Venkatesh Avula) and J.L.; Writing—original draft, V.A. (Vida Abedi) and C.L.; Writing—review & editing, V.A. (Vida Abedi), C.L., D.C., V.A. (Venkatesh Avula), W.H., J.L. and R.Z.; Visualization, V.A. (Vida Abedi), D.C. and V.A. (Venkatesh Avula); Supervision, V.A. (Vida Abedi) and R.Z.; Project administration, V.A. (Vida Abedi) and R.Z.; Funding acquisition, V.A. (Vida Abedi) and R.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Ethical review and approval were waived by the Geisinger Institutional Review Board for this study due to the use of de-identified information.

**Informed Consent Statement:** The study was reviewed and approved by the Geisinger Institutional Review Board to meet “non-human subject research” for using de-identified information.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to institutional policies requiring data-sharing agreement.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Poisson, S.N.; Glidden, D.; Johnston, S.C.; Fullerton, H.J. Deaths from stroke in US young adults, 1989–2009. *Neurology* **2014**, *83*, 2110–2115. [[CrossRef](#)] [[PubMed](#)]
2. Krishnamurthi, R.V.; Moran, A.E.; Feigin, V.L.; Barker-Collo, S.; Norrving, B.; Mensah, G.A.; Taylor, S.; Naghavi, M.; Forouzanfar, M.H.; Nguyen, G.; et al. Stroke Prevalence, Mortality and Disability-Adjusted Life Years in Adults Aged 20–64 Years in 1990–2013: Data from the Global Burden of Disease 2013 Study. *Neuroepidemiology* **2015**, *45*, 190–202. [[CrossRef](#)] [[PubMed](#)]
3. Putaala, J. Ischemic stroke in the young: Current perspectives on incidence, risk factors, and cardiovascular prognosis. *Eur. Stroke J.* **2016**, *1*, 28–40. [[CrossRef](#)] [[PubMed](#)]
4. Lackland, D.T.; Roccella, E.J.; Deutsch, A.F.; Fornage, M.; George, M.G.; Howard, G.; Kissela, B.M.; Kittner, S.J.; Lichtman, J.H.; Lisabeth, L.D.; et al. Factors influencing the decline in stroke mortality a statement from the american heart association/american stroke association. *Stroke* **2014**, *45*, 315–353. [[CrossRef](#)]
5. Hathidara, M.Y.; Saini, V.; Malik, A.M. Stroke in the Young: A Global Update. *Curr. Neurol. Neurosci. Rep.* **2019**, *19*, 91. [[CrossRef](#)]
6. Ji, R.; Schwamm, L.H.; Pervez, M.A.; Singhal, A.B. Ischemic stroke and transient ischemic attack in young adults: Risk factors, diagnostic yield, neuroimaging, and thrombolysis. *Arch. Neurol.* **2013**, *70*, 51–57. [[CrossRef](#)]
7. Jaworek, T.; Xu, H.; Gaynor, B.J.; Cole, J.W.; Rannikmae, K.; Stanne, T.M.; Tomppa, L.; Abedi, V.; Amouyel, P.; Armstrong, N.D.; et al. Contribution of Common Genetic Variants to Risk of Early Onset Ischemic Stroke. *Neurology* **2022**, *99*, e1738–e1754. [[CrossRef](#)]
8. Putaala, J.; Curtze, S.; Hiltunen, S.; Tolppanen, H.; Kaste, M.; Tatlisumak, T. Causes of death and predictors of 5-year mortality in young adults after first-ever ischemic stroke: The Helsinki young stroke registry. *Stroke* **2009**, *40*, 2698–2703. [[CrossRef](#)]
9. Waje-Andreassen, U.; Thomassen, L.; Jusufovic, M.; Power, K.N.; Eide, G.E.; Vedeler, C.A.; Naess, H. Ischaemic stroke at a young age is a serious event—final results of a population-based long-term follow-up in Western Norway. *Eur. J. Neurol.* **2013**, *20*, 818–823. [[CrossRef](#)]

10. Rutten-Jacobs, L.C.A.; Arntz, R.M.; Maaijwee, N.A.M.; Schoonderwaldt, H.C.; Dorresteijn, L.D.; Van Dijk, E.J.; De Leeuw, F.E. Long-term mortality after stroke among adults aged 18 to 50 years. *JAMA* **2013**, *309*, 1136–1144. [[CrossRef](#)]
11. Varona, J.F.; Bermejo, F.; Guerra, J.M.; Molina, J.A. Long-term prognosis of ischemic stroke in young adults: Study of 272 cases. *J. Neurol.* **2004**, *251*, 1507–1514. [[CrossRef](#)] [[PubMed](#)]
12. Béjot, Y.; Daubail, B.; Jacquin, A.; Durier, J.; Osseby, G.V.; Rouaud, O.; Giroud, M. Trends in the incidence of ischaemic stroke in young adults between 1985 and 2011: The dijon stroke registry. *J. Neurol. Neurosurg. Psychiatry* **2014**, *85*, 509–513. [[CrossRef](#)]
13. Cabral, N.L.; Freire, A.T.; Conforto, A.B.; Dos Santos, N.; Reis, F.I.; Nagel, V.; Guesser, V.V.; Safanelli, J.; Longo, A.L. Increase of stroke incidence in young adults in a middle-income country a 10-year population-based study. *Stroke* **2017**, *48*, 2925–2930. [[CrossRef](#)] [[PubMed](#)]
14. Chaudhary, D.; Anyaehie, M.; Demiraj, F.; Bavishi, S.; Shahjouei, S.; Li, J.; Abedi, V.; Zand, R. Comparison of Long-Term Outcomes and Associated Factors between Younger and Older Rural Ischemic Stroke Patients. *J. Clin. Med.* **2022**, *11*, 1430. [[CrossRef](#)] [[PubMed](#)]
15. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Ann. Intern. Med.* **2015**, *162*, W1–W73. [[CrossRef](#)] [[PubMed](#)]
16. Carroll, R.J.; Bastarache, L.; Denny, J.C. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **2014**, *30*, 2375–2376. [[CrossRef](#)]
17. van Alebeek, M.E.; Arntz, R.M.; Ekker, M.S.; Synhaeve, N.E.; AMM Maaijwee, N.; Schoonderwaldt, H.; van der Vlugt, M.J.; van Dijk, E.J.; Rutten-Jacobs, L.C.; de Leeuw, F.-E. Risk factors and mechanisms of stroke in young adults: The FUTURE study. *J. Cereb. Blood Flow Metab.* **2018**, *38*, 1631–1641. [[CrossRef](#)]
18. Sacco, R.L.; Wolf, P.A.; Gorelick, P.B. Risk factors and their management for stroke prevention: Outlook for 1999 and beyond. *Neurology* **1999**, *53*, S15–S24.
19. Kokotailo, R.A.; Hill, M.D. Coding of Stroke and Stroke Risk Factors Using International Classification of Diseases, Revisions 9 and 10. *Stroke* **2005**, *36*, 1776–1781. [[CrossRef](#)]
20. Arboix, A. Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. *World J. Clin. Cases* **2015**, *3*, 418. [[CrossRef](#)]
21. Surtees, P.G.; Wainwright, N.W.J.; Luben, R.N.; Wareham, N.J.; Bingham, S.A.; Khaw, K.T. Psychological distress, major depressive disorder, and risk of stroke. *Neurology* **2008**, *70*, 788–794. [[CrossRef](#)]
22. Lambiase, M.J.; Kubzansky, L.D.; Thurston, R.C. Prospective study of anxiety and incident stroke. *Stroke* **2014**, *45*, 438–443. [[CrossRef](#)] [[PubMed](#)]
23. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [[CrossRef](#)]
24. Abedi, V.; Li, J.; Shivakumar, M.K.; Avula, V.; Chaudhary, D.P.; Shellenberger, M.J.; Khara, H.S.; Zhang, Y.; Lee, M.T.M.; Wolk, D.M.; et al. Increasing the Density of Laboratory Measures for Machine Learning Applications. *J. Clin. Med.* **2020**, *10*, 103. [[CrossRef](#)] [[PubMed](#)]
25. Li, J.; Yan, X.S.; Chaudhary, D.; Avula, V.; Mudiganti, S.; Husby, H.; Shahjouei, S.; Afshar, A.; Stewart, W.F.; Yeasin, M.; et al. Imputation of missing values for electronic health record laboratory data. *NPJ Digit. Med.* **2021**, *4*, 147. [[CrossRef](#)]
26. R Core Team. The R Project for Statistical Computing. 2013. Available online: <http://www.r-project.org/> (accessed on 30 October 2022).
27. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; R Core Team; et al. caret: Classification and Regression Training. R Package Version Vol. 6.0 81. 2015. Available online: <https://CRAN.R-project.org/package=caret> (accessed on 30 October 2022).
28. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Extreme Gradient Boosting. R Package Version 1.1.1.1. 2019. Available online: <https://cran.r-project.org/web/packages/xgboost/index.html> (accessed on 30 October 2022).
29. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
30. YoungStroke—Raising Awareness about Treatment and Management of Stroke in Young Adults. Available online: <https://youngstroke.org/> (accessed on 30 October 2022).
31. Smajlović, D. Strokes in young adults: Epidemiology and prevention. *Vasc. Health Risk Manag.* **2015**, *11*, 157–164. [[CrossRef](#)]
32. Kissela, B.M.; Khoury, J.C.; Alwell, K.; Moomaw, C.J.; Woo, D.; Adeoye, O.; Flaherty, M.L.; Khatri, P.; Ferioli, S.; De Los Rios La Rosa, F.; et al. Age at stroke: Temporal trends in stroke incidence in a large, biracial population. *Neurology* **2012**, *79*, 1781–1787. [[CrossRef](#)]
33. Putaala, J.; Yesilot, N.; Waje-Andreassen, U.; Pitkaniemi, J.; Vassilopoulou, S.; Nardi, K.; Odier, C.; Hofgart, G.; Engelter, S.; Burow, A.; et al. Demographic and geographic vascular risk factor differences in european young adults with ischemic stroke: The 15 cities young stroke study. *Stroke* **2012**, *43*, 2624–2630. [[CrossRef](#)]
34. Siriratnam, P.; Godfrey, A.; O'Connor, E.; Pearce, D.; Hu, C.; Low, A.; Hair, C.; Oqueli, E.; Sharma, A.; Kraemer, T.; et al. Prevalence and risk factors of ischaemic stroke in the young: A regional Australian perspective. *Intern. Med. J.* **2020**, *50*, 698–704. [[CrossRef](#)]
35. Treadwell, S.D.; Robinson, T.G. Cocaine use and stroke. *Postgrad. Med. J.* **2007**, *83*, 389–394. [[CrossRef](#)] [[PubMed](#)]
36. Tuomilehto, J.; Bonita, R.; Stewart, A.; Nissinen, A.; Salonen, J.T. Hypertension, cigarette smoking, and the decline in stroke incidence in eastern Finland. *Stroke* **1991**, *22*, 7–11. [[CrossRef](#)] [[PubMed](#)]

37. Towfighi, A.; Saver, J.L. Stroke declines from third to fourth leading cause of death in the United States: Historical perspective and challenges ahead. *Stroke* **2011**, *42*, 2351–2355. [[CrossRef](#)] [[PubMed](#)]
38. Robinson, R.G.; Jorge, R.E. Post-stroke depression: A review. *Am. J. Psychiatry* **2016**, *173*, 221–231. [[CrossRef](#)]
39. Howard, G.; Kleindorfer, D.O.; Cushman, M.; Long, D.L.; Jasne, A.; Judd, S.E.; Higginbotham, J.C.; Howard, V.J. Contributors to the Excess Stroke Mortality in Rural Areas in the United States. *Stroke* **2017**, *48*, 1773–1778. [[CrossRef](#)]
40. Stamler, J.; Stamler, R.; Neaton, J.D.; Wentworth, D.; Daviglus, M.L.; Garside, D.; Dyer, A.R.; Liu, K.; Greenland, P. Low risk-factor profile and long-term cardiovascular and noncardiovascular mortality and life expectancy. Findings for 5 large cohorts of young adult and middle-aged men and women. *J. Am. Med. Assoc.* **1999**, *282*, 2012–2018. [[CrossRef](#)]
41. Joubert, J.; Prentice, L.F.; Moulin, T.; Liaw, S.-T.S.T.; Joubert, L.B.; Preux, P.-M.P.M.; Ware, D.; De Bustos, E.M.; Mclean, A. Stroke in rural areas and small communities. *Stroke* **2008**, *39*, 1920–1928. [[CrossRef](#)]
42. Lisabeth, L.; Bushnell, C. Menopause and stroke: An epidemiologic review. *Lancet Neurol* **2012**, *11*, 82–91. [[CrossRef](#)]
43. Nichols, H.B.; Trentham-Dietz, A.; Hampton, J.M.; Titus-Ernstoff, L.; Egan, K.M.; Willett, W.C.; Newcomb, P.A. From Menarche to Menopause: Trends among US Women Born from 1912 to 1969. *Am. J. Epidemiol.* **2006**, *164*, 1003–1011. [[CrossRef](#)]
44. Lisabeth, L.D.; Beiser, A.S.; Brown, D.L.; Murabito, J.M.; Kelly-Hayes, M.; Wolf, P.A. Age at natural menopause and risk of ischemic stroke the framingham heart study. *Stroke* **2009**, *40*, 1044–1049. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.