International Journal of Cerebrovascular Disease and Stroke
Brorson JR, et al. Int J Cerebrovasc Dis Stroke 8: 205
www.doi.org/10.29011/2688-8734.100205
www.gavinpublishers.com

Research Article

Creation of a Merged Harmonized Large Database of Subject-Level Data from Acute Secondary Prevention Studies in Minor Non-Cardioembolic Stroke or TIA

James R. Brorson[1*], Simmer Beniwal[2], Mihai Giurcanu[3], Danielle Landron[2], Stacie Landron[2], James E. Siegler[1], Shyam Prabhakaran[1], Julie A. Johnson[2]

[1]Department of Neurology, the University of Chicago, Chicago, IL, USA

[2]Center for Research Informatics, the University of Chicago, Chicago, IL, USA

[3]Department of Health Studies, the University of Chicago, Chicago, IL, USA

***Corresponding author:** James R. Brorson, Department of Neurology, the University of Chicago, Chicago, IL, USA.

Citation: Brorson JR, Beniwal S, Giurcanu M, Landron D, Landron S, et al. (2026) Creation of a Merged Harmonized Large Database of Subject-Level Data from Acute Secondary Prevention Studies in Minor Non-Cardioembolic Stroke or TIA. Int J Cerebrovasc Dis Stroke 9: 205. https://doi.org/10.29011/2688-8734.100205

Received Date: 10 April, 2026; **Accepted Date:** 16 April, 2026; **Published Date:** 20 April, 2026

Abstract

Introduction: Prevention of early recurrence of ischemic stroke after an initial event is a clinically important problem that has been the subject of a number of large clinical trials testing acute treatment interventions. Datasets of clinical trials provide a rich source of information beyond the primary research questions addressed. Merging of datasets, providing greater statistical power and wider applicability for novel exploratory analyses, requires careful harmonization of similar data elements. **Methods:** The Acute Stroke or Transient Ischaemic Attack Treated with Aspirin or Ticagrelor and Patient Outcomes (SOCRATES), Platelet-Oriented Inhibition in New TIA and Minor Ischemic Stroke Trial (POINT), and Acute STroke or Transient IscHaemic Attack Treated with TicAgreLor and ASA for PrEvention of Stroke and Death (THALES) trials each examined the effects of early administration of enhanced antiplatelet therapy versus aspirin alone within 12-24 hours following minor ischemic stroke or transient ischemic attack. **Results:** The datasets of these 3 trials have been harmonized and merged into a single large data table comprising 26319 subjects with 75 clinical and demographic variables, 196 concomitant medication descriptors, and event flags and event timing variables for 20 clinical outcome events. A high degree of data completeness was attained for most variables. **Conclusion:** This merged data table provides a powerful resource of pooled clinical trial data to enable future large-scaled analyses for the study of stroke recurrence risks, and is now curated and available for other investigators.

Keywords: Datasets; stroke recurrence; variables; harmonization; curation.

Introduction

Recurrent stroke commonly follows an initial event, compounding disability and mortality risks. The risk of stroke recurrence is highest in the initial week following a minor non-cardioembolic stroke or TIA. This vulnerable period is a strategic target for secondary stroke prevention. Several major clinical trials, including SOCRATES [1], POINT [2], and THALES [3], have focused on this target with enhanced antiplatelet therapy, initiating treatment acutely within 12-24 hours following the ictus. These 3 trials, all testing addition or substitution of ticagrelor or clopidogrel to aspirin alone, share many similarities in trial structure, design, and core data elements, and show evidence of very similar temporal features of stroke recurrence [4], suggesting comparable clinical populations and risks. These similarities offer the opportunity for subject-level combination of the similar datasets for increased statistical power and generalizability. Combining these large data sets may provide the opportunity to detect which risk factors or medications affect the distinct processes that determine stroke recurrences over time. At the same time, there are some distinctions in definitions of clinical variables that call for careful harmonization in an effort at merging these datasets. Data harmonization, the practice of conforming various sources, types, and structures of data into formats that are compatible and comparable, can provide benefits in statistical power, generalizability, and data quality [5, 6]. As compared to aggregate data meta-analyses, harmonized datasets provide more opportunities for exploring interactions between clinical features and risk modifiers for outcomes of interest that may go beyond the primary outcomes reported in the original trials. Even individual participant-level meta-analysis, while similar in some ways, [7], usually take a narrower variable and outcome focus, and offer less flexibility of analytical questions and ready availability for repurposing and use by external researchers, than does a comprehensive harmonized merged data table, curated and web-accessible. This project describes the approach to harmonization and merging of the 3 large datasets from the SOCRATES, POINT, and THALES trials, and the features of the resulting large merged dataset.

Methods

The studies- similarities and distinctives

The selected studies each targeted subjects with acute non-cardioembolic minor ischemic stroke or high-risk TIA, randomizing subjects within 12-24 hours of index events. Each compared a form of enhanced anti-platelet therapy to treatment with aspirin alone. There were minor differences in the definition of minor stroke in the studies. In SOCRATES and THALES, minor stroke was defined by NIHSS score of 5 or less, while in POINT it was restricted to NIHSS of 3 or less. The ABCD2 score, used to define the risk level of TIA, was required to be 4 or more in SOCRATES and POINT, and 6 or more in THALES. The time window for randomization following the qualifying ischemic event was within 24 hours of onset of symptoms in SOCRATES and THALES, and within 12 hours of the last time known to be free of new symptoms in POINT. The chosen enhanced anti-platelet regimen and the dosing of aspirin differed between the studies, with SOCRATES comparing ticagrelor 90 mg twice daily (following a loading dose of 180 mg) to aspirin 100 mg daily (following a loading dose of 300 mg); POINT comparing clopidogrel 75 mg daily (following loading doses of 600 mg) to placebo, in addition to aspirin 50 mg – 325 mg daily (following a loading dose of 325 mg); and THALES comparing ticagrelor 90 mg twice daily (following 180 mg loading dose) to placebo, added to aspirin 75 mg to 100 mg daily (following loading dose of 300 mg to 325 mg). Each of the trials were randomized with a 1:1 assignment of subjects to treatment groups, and each was double- blinded with placebo controls. SOCRATES and POINT continued study treatment and follow up for 90 days, while THALES continued study medication treatment for 30 days, with subject follow-up for up to 60 days.

Datasets

The POINT trial dataset was provided by the National Institute of Neurological Disease and Stroke, obtained from the NINDS data repository as a set of 22 de-identified SAS-compatible SYSdBat files. The datasets of the SOCRATES and THALES trials, from data contributor AstraZeneca, comprised 23 files and 13 files respectively, and were made available through Vivli, Inc. on a secure server as de-identified SAS-compatible SYSdBat files. Data encoding followed the ADaM (Analysis Data Model) standardized system.

The shared datasets included 4,881 subjects from the POINT trial, 12,018 subjects from the SOCRATES trial, and 9,565 subjects from the THALES trial, totaling 26,464 subjects (Figure 1). Data were not shared for subjects who had not given or had withdrawn consent (113 subjects in SOCRATES and 20 subjects in THALES) nor for subjects from countries not allowing data re-use based on the consent wording (1,182 subjects in SOCRATES and 1,489 subjects in THALES). From the shared datasets, 101 subjects from the SOCRATES trial and 44 subjects from the THALES trial were not assigned to a treatment group and largely lacked any clinical data, and these subjects were excluded from the final merged dataset. Intention-to-treat populations, definitions of outcome events, designation of event times, and data censoring methods were defined according to the trial protocols [1-3]. The POINT dataset served as the default harmonization standard for most clinical and outcome variables, as no participants were excluded from the shared POINT dataset and it was provided with an organized explanatory data dictionary. Data assembly and analysis was carried out using SAS version 9.4 in the Vivli research environment.

Figure 1

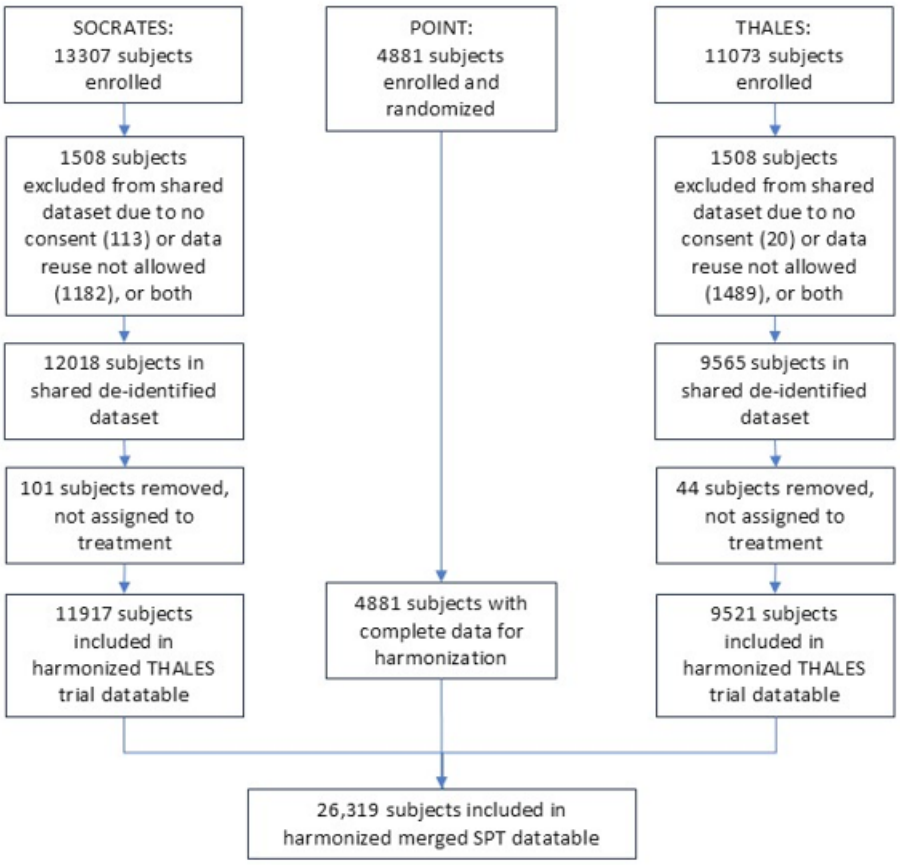

Figure1: Flow diagram of subjects included in the merged dataset.

General Approach to Harmonization

Existing clinical trial datasets were pooled and harmonized through standardized procedures to create a single integrated dataset designed to support future analyses with increased sample size and statistical power. As a foundational step, a target data schema was developed defining standardized variable names, data types, coding conventions, and allowable values. This schema served as the reference framework for mapping and transforming variables from each source dataset, ensuring consistent alignment and integration across the pooled trials. Across the contributing clinical trials, a number of core variables were identified that used broadly similar clinical concepts but differed in naming conventions, coding schemes, measurement, and visit structures. Prior to harmonization, all source datasets were systematically reviewed to identify common variables and assess differences in definitions, formats, and allowable values. While key domains such as demographic characteristics, clinical outcomes, and outcome timing were represented across studies, variability existed in variable labeling, categorical encodings, and measurement structures. These differences informed the development of harmonization rules used to map study-specific variables to the standardized target data schema.

Harmonization strategies - Demographic and clinical data

Clinical data including demographic information, medical history, and baseline evaluations, as recorded in each of the trials, were culled from the trial data sets. Some potentially identifying data such as data of birth, dates of randomization or study visits, and country of origin were excluded in the de-identification process by the data owners prior to provision in all 3 datasets. Specific subject ages were not provided in the SOCRATES and THALES trials; instead, age ranges were specified. Specific ages (in years) up to 89 were provided in POINT. Body mass index (BMI) values were only provided as ranges in SOCRATES and THALES, and were

not provided in the POINT dataset.

Unknown data values were of two types. Certain variables were simply not recorded in a given trial. Examples include smoking history, not included in SOCRATES, or history of prior stroke, not included in POINT. Such variable values were considered 'not reported' and in the harmonization process were designated as such by blank values for character variables and by entry of '-999' for numeric variables (known values for these variables are all non-negative). Other variable values were within the scope of a study's protocol but not found; these were designated with the entry '.' to indicate 'missing' data values.

Specific flags for certain medical history conditions, such as history of hypertension, history of diabetes, were recorded in all 3 trials. Other medical history conditions were discoverable in SOCRATES and THALES trials by search of the provided medical history files for listing of the specific condition name; if not found listed for a subject, the history of that medical condition was considered not present and in harmonization the corresponding flag was imputed a value of 'N'.

Harmonization of Medication data

All 3 trials included flags for aspirin exposure immediately prior to randomization. Prior exposures to certain other antiplatelet agents, to statins, and to other lipid-lowering medications were indicated by specific flags in POINT and were available in the concomitant medication data files for SOCRATES and THALES. The POINT trial also provided a limited information about exposure on the dates of each study visit to statin medications, to NSAIDS, to oral anticoagulant medications, and to proton pump inhibitors, but did not provide start and end dates regarding these medications, and did not report other concomitant medication use. SOCRATES and THALES in contrast included detailed concomitant medication data files with comprehensive reporting of all medications used during the trial by each subject, with specific medication names and Anatomic Therapeutic Chemical (ATC) codes, start and end dates, and dosages.

Twenty medications or medication groups of interest were selected for inclusion in the harmonized merged database. Searches for these concomitant medication exposures in SOCRATES and THALES datasets were generally based on ATC codes, with certain exceptions (see Table 1). In certain cases, medication names were used. For example, cilostazol use was pulled by name, to separate it from other antiplatelets included under the same ATC code. For prior antiplatelet use variables, searches were by medication names.

Variable	Description	Search term
PRASA	Aspirin use in the 7 days prior to randomization	Specific flags for prior aspirin use available in all 3 studies
PRCLO	Clopidogrel use prior to randomization	Specific flag in POINT; search by names 'CLOPIDOGREL' or "CLOPIDOGREL SULFATE' in SOCRATES and THALES
PRDIP	Dipyridamole use prior to randomization	Specific flag in POINT; search by names 'DIPYRIDAMOLE' or 'ACETYLSALICYLIC ACID+DIPYRIDAMOLE' in SOCRATES and THALES
PRTIC	Ticlopidine use prior to randomization	Specific flag in POINT; search by name 'TICLOPIDINE' in SOCRATES and THALES
PROAP	Other anti-platelet use prior to randomization, excluding aspirin, clopidogrel, dipyridamole, ticlopidine, and cilostazol.	ATC code B01AC, excluding aspirin, clopidogrel, dipyridamole, ticlopidine, and cilostazol.
STA**	Statin medication exposure	ATC codes C10AA, C10BA, and C10BX
LLM**	Other lipid-lowering medications exposure	ATC codes C10AB, C10AC, C10AD, C10AX
NSD**	Non-steroidal anti-inflammatory medications exposure	ATC codes M01AA, M1AB, M01AC, M01AE, M01AG, and M01AH
CIL**	Cilostazol exposure	Medication name 'CILOSTAZOL'
OAC**	Oral anticoagulant medication exposure	ATC code B01AA, B01AE, or B01AF
HEP**	Heparin or heparinoid medication exposure	ATC code B01AB
PPI**	Proton pump inhibitor medication exposure	ATC code A02BC
STE**	Steroid medication exposure	ATC code H02AB orH02BX

COL**	Colchicine exposure	ATC code M04AC
OIF**	Other anti-inflammatory medication exposure	ATC code L04AA, L04AX, or M01AX
NOO**	Nootropic medication exposure	ATC code N06BX
AIB**	ACE inhibitor or angiotensin receptor blocker medication exposure	ATC code C09AA, C09BA, C09BB, C09BX, C09CA, C09DA, C09DB, or C09DX
BBL**	Beta-blocker medication exposure	ATC code C07AA, C07AB, C07AG, C07BA, C07BB, C07CA, C07CB, C07DA, C07DB, C07EA, C07EB, C07FB, or C07FX
CCB**	Ca channel blocker medication exposure	ATC codes C08CA, C08CX, C08DA, C08DB, C08EA, C08EX, C08GA, or C07FB
THZ**	Thiazide diuretic medication exposure	ATC codes C03AA, C03AB, C03AH, C03 AX, C07BA, C07BB, or C07BG
HYP**	Other antihypertensive medications exposure	ATC codes C02AB, C02AC, C02CA, C02DA, C02DB, C02DC, C02DD, C02DG, C03BA, C03DA, C03DB, or C09XA
MET**	Metformin/biguanide medication exposure	ATC code A10BA or A10BD
SUL**	Sulfonylurea medication exposure	ATC code A10BB or A10BD
GLP**	GLP-1 agonist medication exposure	ATC code A10BJ

Table 1: Search and data summarizing strategies for concomitant medications. Full variable names are given for flags for prior antiplatelet medication use. For other medication variables, a 3 letter code designates the medication group, and a 2 letter suffix defines the parameter pertaining to that medication that the variable describes.

Several challenges arose in characterizing the detailed concomitant medication exposure information in a data table format with a single row assigned for each subject. Challenges included multiple exposures to different medications of the same class (i.e., different statins), multiple short-term exposures with different start and stop dates, overlapping exposure periods, and missing data for start and stop dates. This complex information was condensed into a set of 10 harmonized variables for each medication group summarizing the exposures, including start and end days, number of exposures, and total exposure duration. Details of the medication -related variables and of the naming conventions followed can be found in the explanatory READ ME file included in Supplementary materials.

Harmonization of Outcome Events

A wide variety of pre-specified outcome events was available for each study. Flags for each study's primary outcome event, which was defined differently for the 3 studies (any stroke, myocardial infarction, or death for SOCRATES; ischemic stroke, myocardial infarction, or ischemic vascular death for POINT; and any stroke or death for THALES), were mapped to a single merged primary outcome event flag. Ischemic stroke, the most commonly occurring outcome event, was similarly defined in each trial and mapped to a single event flag. Multiple other important outcomes were tracked in all 3 trials, though with some minor differences of definition, and were mapped to common harmonized event flags, including flags for hemorrhagic stroke, any stroke, death, and major hemorrhage events. A total of 20 different outcome events common to at least 2 of the trials were assigned to harmonized event flags and event time variables, following the convention of the SOCRATES and THALES trials with failure events coded as '0' and censored subjects assigned '1'. The harmonized timing flags were assigned values of the nearest number of days elapsed from day of randomization, which was assigned a value of '0'. This required rounding to the nearest whole day of the timing variable in the POINT trial, which were recorded as decimal numbers of days, and subtraction of 1 day in the timing variables of outcomes in the SOCRATES and THALES trial datasets, which followed a convention calling the day of randomization as day 1.

Further details of harmonization conventions followed and variable naming conventions are described in the README file that is included in supplementary materials and on the Vivli.org platform. Full descriptions of each variable and how the harmonized variable relates to the source data of the individual trial datasets are found in the SPT Data Dictionary, also included in the supplementary materials and on the Vivli platform.

Results

Harmonization of demographic and clinical variables

The harmonized merged data table includes a total of 26,319 subjects, comprising 90% of the total numbers of subjects enrolled in the three trials. A total of 311 variables were included, as described in Table 2.

Section	Variable groups included
Demographic and Clinical Variables (75 variables)	Header (3), Demographics (4), Randomization data (13), Baseline clinical data (25), Medical history (18), End of study data (12)
Medication Variables (20 medication groups, with 6 medication parameters for the antiplatelet medication group, 10 for others)	Prior antiplatelets, Statins, Other lipid-lowering medications, NSAIDs, Cilostazol, Oral anticoagulants, Heparinoids, Proton pump inhibitors, Steroid medications, Colchicine, Other anti-inflammatory medications, Nootropic medications, ACE inhibitors or ARBs, Beta-blockers, Ca channel blockers, Thiazide diuretics, Other antihypertensives, Metformin, Sulfonylureas, Glucagon-like peptide agonists.
Outcome Event parameters (20 outcome events, each with a censoring flag variable and an event timing variable) (ITT= intention-to-treat, AT= As-treated)	Composite primary outcome – ITT, Ischemic stroke – ITT, hemorrhagic stroke- ITT, TIA-ITT, Vascular Death -ITT, Any Death – ITT, Major Hemorrhage – ITT, Ischemic vascular death -ITT, MI – ITT, Symptomatic ICH – ITT, Fatal Hemorrhage – ITT, Composite primary outcome – AT, Major hemorrhage – AT, Fatal hemorrhage – AT, Intracranial hemorrhage – AT, Ischemic stroke – AT, Hemorrhagic stroke – AT, Any stroke – AT, Any Death – AT

Table 2: Summary of variables included in SPT datatable: Detailed descriptions of all included variables are available in the SPT Data Dictionary, included among the supplementary materials

A high level of data completeness was achieved, with numbers of subjects with known variable values exceeding 80% for 55 of 75 demographic and clinical variables, for 30 of 40 outcome event variables, and for 111 of 196 concomitant medication parameters (Figure 2, and Table 3). Concomitant medication data were not collected in the POINT trial (save for antiplatelet agents and statins), preventing a high level of >80% completeness for most of these variables. Other notable gaps in collected data included lack of baseline laboratory data in the SOCRATES and THALES datasets, lack of smoking status and of carotid or intracranial artery stenosis data in the SOCRATES trial dataset, and lack of end-of-study vital signs in POINT.

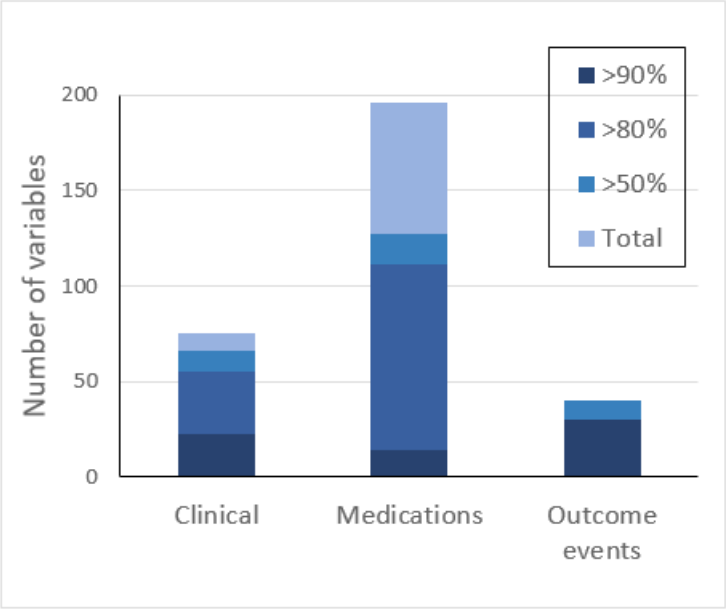

Figure 2: Data completeness. Numbers of included variables, and variable completeness at 50%, 80%, and 90% thresholds.

	Demographic and Clinical variables (N=75)	Concomitant Medication parameters (20 medication groups, 6 - 10 parameters per medication group*)	Outcome Events, censoring flag and event times (20 outcome events, 2 variables per OE)
$N_{50\%}$	66	127	40
$N_{80\%}$	55	111	30
$N_{90\%}$	23	14	30
Total no. of variables	75	196	40

Table 3: Data Completeness metrics – Numbers of variables with known values above 50%, 80%, and 90% thresholds

Concomitant medication usage

Exploratory studies of effects of concomitant medications on stroke-related outcomes are of great interest. Such analyses will require adequate exposures of subjects to the medications in question. Exposures (based on the concomitant use variable) varied considerably across medication groups (Table 4). As expected, most patients (64%) in the trials were treated with statin medications but only a small minority were treated with other lipid-lowering medications. Most patients were on at least one antihypertensive medication, with angiotensin -converting enzyme (ACE) inhibitors or angiotensin receptor blocker (ARB) medications, beta-blocking medications, and Calcium channel-blocking medications being the most common. Many were on medications for diabetes, but glucagon-like peptide-1 (GLP-1) agonist use was rare during these trials. A number of subjects were exposed to NSAIDs, but few were treated with other anti-inflammatory medications. A large number of subjects were concurrently treated with nootropic medications.

Medication group	Known (%)	Not reported	Missing	Number of exposed subjects
Statins	26319 (100%)	0	0	20842
Other lipid-lowering medications	21438 (81.5%)	4881	0	846
NSAIDs	26319 (100%)	0	0	2061
Cilostazol	21438 (81.5%)	4881	0	171
Oral anticoagulants	26319 (100%)	0	0	1400
heparinoids	21438 (81.5%)	4881	0	2175
PPIs	26319 (100%)	0	0	9570
Steroid medications	21438 (81.5%)	4881	0	690
Colchicine	21438 (81.5%)	4881	0	170
Other anti-inflammatory medications	21438 (81.5%)	4881	0	250
Nootropic medications	21438 (81.5%)	4881	0	6354
ACE inhibitors/ARBs	21438 (81.5%)	4881	0	14244
Beta-blocking medications	21438 (81.5%)	4881	0	6027
Ca-channel blocking medications	21438 (81.5%)	4881	0	7098
Thiazide diuretics	21438 (81.5%)	4881	0	1388
Other antihypertensives	21438 (81.5%)	4881	0	3280
Metformin	21438 (81.5%)	4881	0	4149
Sulfonylurea medications	21438 (81.5%)	4881	0	2252
GLP-1 analogue medications	21438 (81.5%)	4881	0	30

Table 4: Concomitant medication use: Completeness and exposure levels for the different medication groups for subjects included in the SPT merged datatable

Outcome events: completeness and event numbers

Most outcome events were included in data recording in all 3 trials, resulting in data that are complete or near-complete for most outcomes of interest. All 3 trial data sets tracked most of the same outcome events of interest including stroke subtypes, both ischemic and hemorrhagic, major hemorrhages, and fatal events. The THALES trial dataset did not report outcomes of 'Any Stroke', TIA', "MI', or 'Vascular death', so these outcomes are not reported for the 9521 subjects from this trial. On-treatment status was not flagged in 108 of the 4881 subjects in the POINT trial, so that the As-treated outcomes for these subjects are recorded as missing data. Event numbers, crucial for statistical power estimation for any proposed analyses, were moderately high for outcomes that included ischemic stroke, but were only modest or low for most other outcomes of interest. (See Table 5).

Event type	Included subjects (% of total 26,319)	Not reported	Missing	Number of events
Composite primary outcome - ITT	26319 (100%)	0	0	1529
Composite primary outcome - AT	26211 (99.6%)	0	108	1462
Ischemic stroke outcome - ITT	26319 (100%)	0	0	1380
Ischemic stroke outcome - AT	26211 (99.6%)	0	108	1369
Hemorrhagic stroke outcome - ITT	26319 (100%)	0	0	29
Hemorrhagic stroke outcome - AT	26211 (99.6%)	0	108	26
Any stroke - ITT	16798 (63.8%)	9521*	0	970
Any stroke - AT	26319 (100%)	0	108	1391
Transient ischemic attack- ITT	16798 (63.8%)	9521*	0	428
Vascular Death - ITT	16798 (63.8%)	9521*	0	86
Ischemic vascular death - ITT	16798 (63.8%)	9521*	0	67
Any death – ITT	26319 (100%)	0	0	194
Any Death – as treated	26211 (99.6%)	0	108	141
Myocardial infarction - ITT	16798 (63.8%)	9521*	0	58
Major Hemorrhage - ITT	26319 (100%)	0	0	130
Major hemorrhage - AT	26211 (99.6%)	0	108	118
Symptomatic intracranial hemorrhage - ITT	26319 (100%)	0	0	61
Fatal hemorrhage - ITT	26319 (100%)	0	0	29
Fatal hemorrhage - AT	26211 (99.6%)	0	108	22
Intracranial hemorrhage - AT	26211 (99.6%)	0	108	52
*Data not provided in the THALES dataset.				

Table 5: Outcome events included, with incompleteness and event numbers (ITT- intention to treat; AT-as treated)

Data curation and availability.

The harmonized merged dataset, as well as an explanatory README file, the code used for its production, the individual trial harmonized data tables, metadata, and an extensively annotated data dictionary, are all housed on the Vivli organization (vivli.org) platform, and are available to qualified investigative teams for analysis. The platform allows users access to common software tools including SAS, R, and Excel, and provides for uploading of users' code into the platform. Access is available through application to Vivli as vetted by the Vivli Independent Review Panel and by approval of the data provider (for SOCRATES and THALES) Astra-Zeneca.

Discussion

Data of major randomized controlled trials provide a rich source of information for exploratory studies testing plausibility and feasibility of additional hypotheses that go beyond the outcome objectives of the original studies. Power for such exploratory studies is enhanced by the combination of parallel trial datasets that include similar clinical groups. The SOCRATES, POINT, and THALES trials addressed very similar clinical patient groups of subjects with acute minor ischemic stroke or TIA, and with similar interventions of early treatment with enhanced antiplatelet therapy, being compared to a control group on aspirin alone.

Several previous efforts have involved harmonization and merging of stroke-related study data. The Risk of Paradoxical Embolism (RoPE) Study focused on databases that included subjects with cryptogenic stroke with investigation for patent foramen ovale, and was able to harmonize and assemble the data from 12 component databases including 3674 subjects aiming to facilitate risk modeling of PFO-attributable stroke recurrence risk [8]. This harmonized database was successful in producing the familiar RoPE risk assessment score validated as identifying groups of cryptogenic stroke patients with heterogeneous risk-related factors suggesting different stroke mechanisms [8, 9]. Its focus was on a narrower group of stroke patients and on a longer period of risk assessment as compared to the broad non-cardioembolic ischemic stroke patient group and immediate post-stroke monitoring period of the present SPT database. A more recent analysis pooled individual subject data from 9 ischemic stroke cohort studies including 1568 subjects that focused on brain MRI imaging and poststroke cognitive assessment [10], successfully identifying an association of white matter hyperintensity volume with decline in cognitive functioning after stroke, but again not providing as broad of a population of included stroke patients nor of related clinical variables as the present database. Finally, Mallya et al. [11] provided a harmonized metadata repository comprising 4 different population-based studies of stroke risk factors. This platform offers users access to harmonized mapping of variables, facilitating examination of the resulting metadata with a flexible interactive online portal for exploring questions in this population. User access is available through the American Heart Association Precision Medison Platform. The present harmonized merged dataset differs in a number of ways from previously described datasets, particularly in the focus on a broad population of subjects with acute minor stroke or TIA, in looking at event timing in the acute period immediately following the ischemic event, in providing a broad set of variables covering demographic, clinical concomitant medication, and outcome event features, and in providing user direct access to this subject-level data.

Conclusions

The harmonized and merged data from these trials is now available for interested investigators, with an extensive accompanying array of demographic, clinical, medication, and outcome variables, opening new opportunities to explore a number of research questions of interest regarding this patient population.

Acknowledgement

This manuscript is based on research using data from data contributor AstraZeneca that have been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

Statement of Ethics

The present study was submitted to the University of Chicago Institutional Review Board as IRB20-1882 and was determined to be exempt from a requirement for written informed consent and from further review.

Conflict of Interest Statement

This project was supported by NIH awards R61 NS135583 and R33NS135583, “Kinetic analysis of acute stroke secondary prevention trials: Insights from combined datasets guiding future trial design” (JRB). The funder had no role in the design, data collection, data analysis, and reporting of this study. There are no other competing interests of relevance to the present work.

Author Contributions

JRB conceived of the project and contributed to organization, coding, data harmonization, and drafting the manuscript. SB contributed to data harmonization and coding. MH provided statistical review. DL and SL assisted with data harmonization. SP and JWS provided manuscript review and editing. JJ assisted with methodologies, data harmonization, and manuscript review.

Data Availability Statement

The individual study data tables and the harmonized merged data table described in this manuscript, as well as coding files used to generate the data tables, the Data Dictionary, a spreadsheet of data harmonization metrics, and an explanatory READ ME file, are all available to interested parties on the Vivli.org platform through application to Vivli.

References

1. Johnston SC, Amarenco P, Albers GW, Denison H, Easton JD, et al. (2016) Ticagrelor versus aspirin in acute ischemic stroke or transient ischemic attack. N. Engl. J. Med. 375: 35–43.
2. Johnston SC, Easton JD, Farrant M, Barsan W, Conwit RA, et al. (2018) Clopidogrel and aspirin in acute ischemic stroke and high-risk TIA. N Engl J Med. 379: 215-225.
3. Johnston SC, Amarenco P, Denison H, Evans SR, Himmelmann A, et al. (2020) Ticagrelor and aspirin or aspirin alone in acute ischemic stroke or TIA. N Engl J Med. 383: 207-217.
4. Brorson JR, Giurcanu M, Prabhakaran S, Johnston SC (2023) Vulnerable and stabilized states after cerebral ischemic event: Implications of kinetic modeling in the SOCRATES, POINT, and THALES trials. Neurology. 101: e2205-e2214.

5. Cheng C, Messerschmidt L, Bravo I, Waldbauer M, Bhavikatti R, et al. (2024) A general primer for data harmonization. Scientific Data. 11: 152.

6. Adhikari K, Patten SB, Patel AB, Premji S, Tough S, et al. (2021) Data harmonization and data pooling from cohort studies: a practical approach for data management. Int J Popul Data Sci. 6: 1680.

7. Veroniki AA, Seitidis G, Tsivgoulis G, Katsanos AH, Mvridis D (2023) An Introduction to Individual Participant Data Meta-analysis. Neurology. 100:1102-1110.

8. Thaler DE, Di Angelantonio E, Di Tullio MR, Donovan JS, Griffith J, et al. (2013) The risk of paradoxical embolism (RoPE) study: Initial description of the completed database. Int J Stroke. 8: 612-619.

9. Thaler DE, Ruthazer R, Weimar C, Mas JL, Serena J, et al. (2014) Recurrent stroke predictors differ in medically treated patients with pathogenic vs other PFOs. Neurology. 83: 221-226.

10. De Kort FAS, Coenen M, Weaver NA, Kuijf HJ, Aben HP, et al. (2023) White matter hyperintensity volume and poststroke cognition: An individual patient data pooled analysis of 9 ischemic stroke cohort studies. Stroke. 54: 3021-3029.

11. Mallya P, Stevens LM, Zhao J, Hong C, Henao R, et al. (2023) Facilitating harmonization of variables in Framingham, MESA, ARIC, and REGARDS studies through a metadata repository. Circ. Cardiovasc Qual Outcomes. 16: e009938.