Integration of Regional Hospitalizations, Registry and Vital Statistics Data for Development of a Single Statewide Ischemic Stroke Database

Zhiyu Yan, MS,[a] Victoria Nielsen, MPH,[b] Glory Song, MPH,[b] Anita Christie, RN, MHA,[b] Lee H. Schwamm, MD,[a] and Kori S. Zachrison, MD, MSc[c]

Objective: Administrative databases seldom include detailed clinical variables and vital status, limiting the scope of population-based studies. We demonstrate a comprehensive process for integrating 3 databases (all-payor inpatient hospitalizations, clinical acute stroke registry and vital statistics) into a single statewide ischemic stroke database. *Materials and methods:* The 3 Massachusetts databases spanned 2007-2017. Our integration process was composed of 3 phases: 1) hospitalizations-registry linkage, 2) hospitalizations-vital linkage, and 3) final integration of all 3 databases. Following data uniqueness assessment, rule-based deterministic linkage on indirect identifiers were applied in the first two phases. We validated the linkages by comparing additional patient variables not used in the linkage process in the absence of a gold standard database crosswalk. *Results:* During the overlapping period from 1/1/2008 to 9/30/2015, there were 47,713 stroke admissions in the hospitalizations database and 43,487 admissions in the registry. We linked 38,493 (80.7%) of cases, 95% of which were validated. There were 391,176 deaths reported in Massachusetts between 1/1/2010 and 3/6/2017 in the vital database. Of the 38,493 encounters in the hospitalizations-registry linked data, 10,660 (27.7%) were linked to deaths, reflecting the cumulative mortality over the 7-year period among all registry-linked ischemic stroke hospitalization records. *Conclusion:* We demonstrate that a high-quality integration of the statewide hospitalizations, clinical registry, and vital statistics databases is achievable leveraging indirect identifiers. This data integration framework takes advantage of rich clinical data in registries and long term outcomes from hospitalizations and vital records and may have value for larger scale outcomes research.

Key Words: Ischemic stroke—Health services research—Database integration—Clinical registry—Vital statistics

Introduction

Ischemic stroke is an important cause of long-term disability[1] and mortality.[2] Thus, longitudinal post-discharge data is valuable for stroke population and health services research. Administrative hospitalization records may provide such information; however, many clinical and population studies require more granular clinical details such as the data included in clinical registries. Additionally, for the consideration of longer-term mortality outcomes, vital records are a critical source of data. The integration of hospitalizations, registry and vital data could harness the relative advantages of each data source into a combined,

From the [a]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States; [b]Office of Statistics and Evaluation, Massachusetts Department of Public Health, Boston, MA, United States; and [c]Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, United States.

Received July 14, 2021; revision received October 29, 2021; accepted November 20, 2021.

Corresponding author at: Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, United States E-mail: zyan6@mgh.harvard.edu.

1052-3057/$ - see front matter

https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106236

comprehensive database. For example, one could investigate the comparative effectiveness of an endovascular intervention (captured by a clinical registry) during the acute stroke episode on patients' long-term outcomes with the integrated data. This would involve modeling post-stroke comorbidities and healthcare resource utilization (captured by administrative data), as well as different modes or causes of mortality (captured by vital records) while adjusting for various prior medications and comorbidities (captured by registry and administrative data). As unique patient identifiers across these three types of databases typically do not exist, indirect identifiers such as patient gender, date of birth and date of service are often used to link data. While previous work has demonstrated the process of linking either registry and hospitalizations data[3,4] or hospitalizations data and vital statistics,[5,6] work that demonstrates a systematic integration of hospitalizations, registry, and vital data is sparse. This may be due to challenges in finding combinations of indirect identifiers to identify records in all three types of databases uniquely. Additionally, the different structures of the individual databases may complicate the integration process.

We aimed to address this gap in the literature by demonstrating a comprehensive process for integrating three databases (all-payor inpatient hospitalizations, stroke clinical registry, and vital statistics) into a single statewide ischemic stroke database using a series of deterministic linkage strategies. With the integrated database, an investigator can then track healthcare utilization and patient mortality status subsequent to the index stroke hospitalization. We first introduced a method for formally assessing the feasibility of linkage based on the calculation of uniqueness among values of indirect patient identifiers. We also demonstrated transforming the structure of hospitalizations data to accommodate its linkage with both registry and vital records. Finally, we demonstrated a process to evaluate linkage quality in the absence of gold standard database crosswalks. This integration framework may serve as a reference for similar three-way database linkages for other geographical regions and clinical conditions.

Methods and results

Data sources and population

Three distinct databases with different timeframes were employed in this study in sequence: the hospitalizations, registry, and vital statistics databases. First, we used hospitalizations records from the Massachusetts (MA) Case Mix Hospital Inpatient Discharge Data[7] (hospitalizations), a comprehensive administrative database that includes all emergency department, outpatient observation, and inpatient discharges from acute care hospitals in Massachusetts. As such, it comprehensively covers all stroke cases requiring emergency treatment in Massachusetts hospitals. Case Mix data also includes detailed demographics, treatments administered, length of hospital stay, comorbidities, total charges, and procedures. It does not include data from hospitals operated by the Veterans Administration, rehabilitation facilities, or ambulatory providers. For this analysis, we included all 69,282 hospitalizations with the primary diagnosis of ischemic stroke (ICD-9 codes: 433.xx excluding 433.10, 434.xx, 436.xx), valid patient identifiers, and discharge date between 10/1/2007 and 9/30/2015. We chose to link on index stroke hospitalizations because the database contains a unique patient identification number that can be used to identify patients' subsequent emergency department visits, observation stays, admissions, and other Case Mix records.

Second, we added the registry records from the MA Paul Coverdell stroke registry[8] (registry). The Coverdell registry is a detailed clinical database that includes all ischemic stroke encounters at participating MA hospitals. For this analysis, we included all 43,487 hospitalizations associated with the primary diagnosis of ischemic stroke and the discharge date between 1/1/2008 and 9/30/2015, during which 49 MA hospitals were participating in the registry. Although this does not include all MA hospitals, the registry does capture approximately 61% of all ischemic stroke encounters in the state.

Third, we added the death records from the MA Vital Statistics data[9] (vital), which includes identifiable mortality data on all deaths in MA. The death records provide information on patients' occupation, death place, death manner, and place of injury when applicable. We were approved to have access to 391,176 records corresponding to all deaths that occurred between 1/1/2010 and 3/6/2017. The key metrics of our data sources are summarized in Table 1. This analysis was approved by the Massachusetts Department of Public Health Institutional Review Board.

Phased linkage process

The overall method can be broken down into a series of deterministic linkage strategies followed by a final integration of three distinct databases into a consolidated

Table 1. *Key metrics of source databases.*

	Time period	Observation level	Number of observations	Reporting hospitals
Hospitalizations	10/1/2007 – 9/30/2015	Encounter	69,282	All 69 MA acute care hospitals
Registry	1/1/2008 – 9/30/2015	Encounter	43,487	49 Coverdell participating hospitals
Vital	1/1/2010 – 3/6/2017	Individual	391,176	N/A

Table 2. *Data populations for each integration phase.*

Phase	Population	Number of Candidate Records to be Linked (% of Original Records)
Hospitalizations-Registry Linkage	Ischemic stroke hospitalizations of patients discharged between 1/1/2008 and 9/30/2015 from the 49 Coverdell-participating hospitals	47,713 (69%) hospitalizations records and 43,487 (100%) registry records
Hospitalizations-Vital Linkage	Ischemic stroke hospitalizations* of patients who were discharged between 10/1/2007 and 9/30/2015 from all 69 MA acute care hospitals, and died between 1/1/2010 and 3/6/2017	69,282 (100%) hospitalizations records and 391,176 (100%) vital records
Integration of the three databases	Ischemic stroke hospitalizations of patients discharged between 1/1/2008 and 9/30/2015 from the 49 Coverdell-participating hospitals	47,713 (69%) hospitalizations records, 43,487 (100%) registry records and 391,176 (100%) vital records

*When an encounter-level database joins with an individual-level database, the resulting linked database is at the encounter-level as one patient may have multiple hospitalizations encounters, but not vice versa.

statewide stroke database. The hospitalizations-registry-vital integration can be divided into three phases: hospitalizations-registry linkage, hospitalizations-vital linkage, and integration of the three databases. Given three source databases, any two out of the three distinct two-way linkages can be implemented during the first two phases. However, some linkages can be superior to others when we consider the choice of master data file for the process. Master data file refers to the data file that initiates searches during a linkage.[10] For example, defining the hospitalizations data as the master means that we would identify an encounter in the hospitalizations database and then search for a matched record in registry or vital databases. It is convenient to have the master data involved in both linkages of choices so that at the very last integration phase, the record identification of the master data can be directly used to join the linked data files obtained from the previous two phases. For our analysis, we chose the hospitalizations database as the master because i) it has the same data level – hospitalization encounter – as the desired integrated data, and ii) it has more records and a longer duration of data than the registry (this is preferable in order to link vital records to as many hospitalizations as possible), as shown in Table 1. It follows that the first two phases correspond to the hospitalizations-registry and hospitalizations-vital linkages.

Decomposing the process into these distinct phases allows us to account for situations when databases correspond to overlapping but not identical populations. The population of the hospitalizations-registry linkage phase is defined as the intersection of the populations of the hospitalizations and registry databases because the encounters of the hospitals not participating in the Coverdell program and the encounters prior to 1/1/2008 are not captured by the registry database. The population of the hospitalizations-vital linkage is defined as the entire hospitalizations population because all patients in the hospitalizations are under the risk of death. Finally, the population of the final integration is defined as the intersection between the populations of the hospitalizations-registry linkage and hospitalization-vital linkage phases. This approach enables the linked data files obtained in each phase to retain their respective maximum populations. When defining phase-specific populations, we need to account for all the key metrics summarized in Table 1. The targeted phase-specific population can then be used to identify candidate records to be linked at that phase, as shown in Table 2. Figure 1 illustrates these three phases in the full integration process.

Hospitalizations-registry linkage

As a potential linkage across databases is based on their common identifiers, the first step in the linkage is to assess whether these identifiers contain sufficient information to support the linkage. If a combination of identifier values cannot uniquely identify most of the records in a source database, then the subsequent linkage may not be feasible. Evaluating data uniqueness informs our assessment of feasibility. Data uniqueness refers to the proportion of the records that can be uniquely identified with a combination of identifier values in a database. For example, when using distinct values of hospital name, admission date, and patient gender as the combination, a record is considered unique if no other record has the same combination of values on all these identifiers. If there is only one female in the database admitted to *Hospital i* on 1/1/2011, this is considered a unique record. The data uniqueness is a summary statistic, giving the proportion of such unique records among all records in a given database. A combination resulting in high uniqueness within both databases involved in the linkage is necessary for the linkage to be feasible. We examine uniqueness separately in each database that will be linked to ensure feasibility.

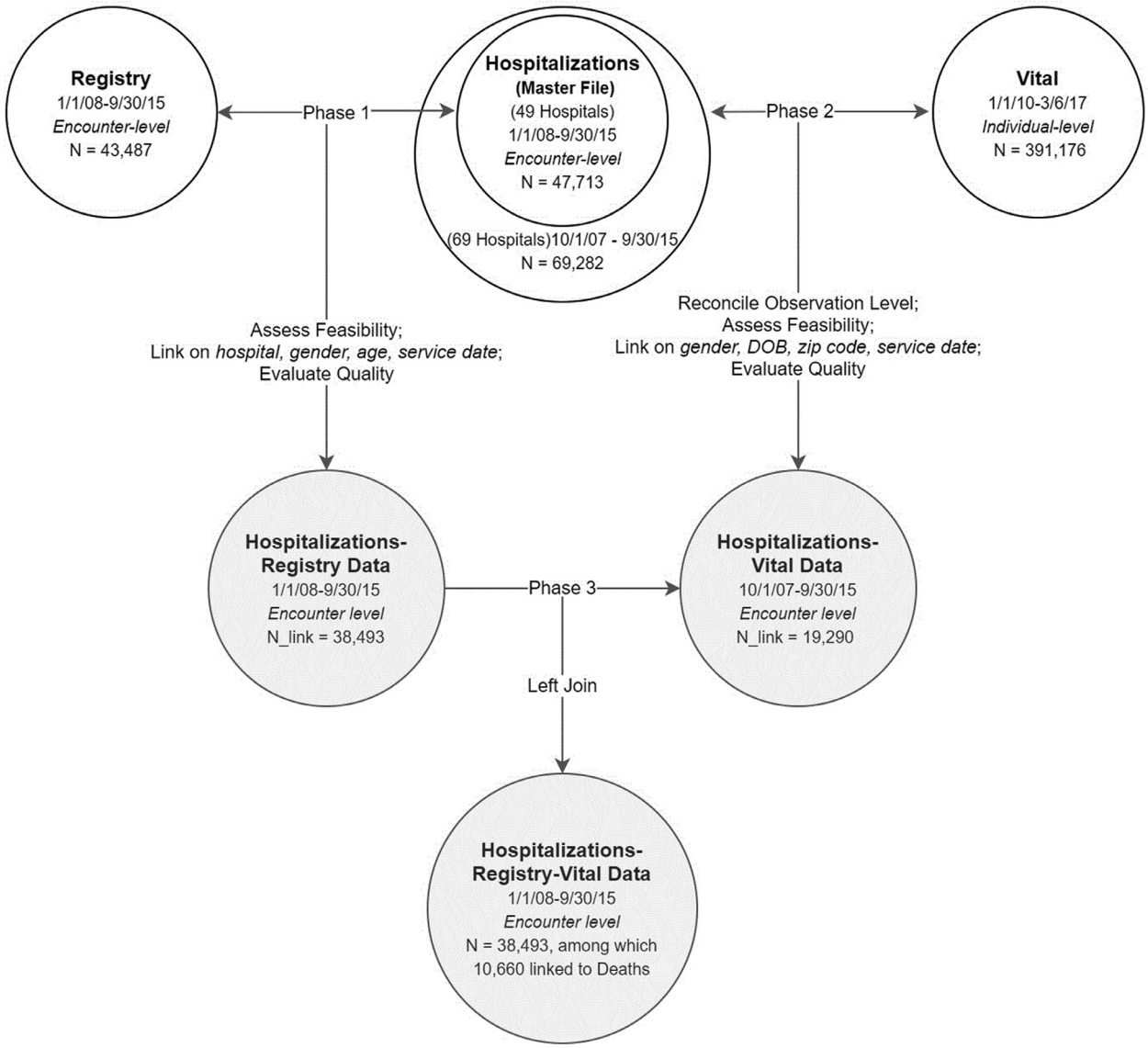

Fig. 1. *Illustration of database populations and integration phases. Legend: N_link: The number of linked records during the step.*

We identified five highly reliable common identifiers in the hospitalizations and registry databases: hospital, patient age, gender, admission date, and discharge date. The combination of these five variables on their distinct values distinguishes over 99% of the records in both hospitalizations and registry databases, supporting that the linkage is feasible with these identifiers. We then checked uniqueness corresponding to the combination of the five identifiers when allowing flexibility in their values. This process involves permitting pre-specified variability in select identifiers (e.g., allowing patient age to disagree by 1 year) in order to account for unknown timing of birth dates relative to hospitalizations, or other similarly minor inconsistencies in the data. Table 3 illustrates the specified points of variability that we examined. All 4 combinations of identifier values with permissive variability as specified still retain relatively high uniqueness (over 97%), and thus can be used to define rules in the deterministic linkage of the two databases.

Following a previously described method,[3] when performing the hospitalizations-vital linkage, we applied all 4 rules shown in Table 3 successively. Rules with higher uniqueness are more likely to give a higher proportion of one-to-one linked pairs, and therefore were assigned a higher priority when applied to linkage. Applying the rules in series, records left unlinked with a given rule became candidates to be linked with the rule of the next lower uniqueness level, as shown in Table 4. When applying each rule, we only retain one-to-one links. After exhausting all 4 rules, we reached 38,493 linked record pairs, which corresponds to 88.5% of the 43,487 candidate registry records and 80.7% of the 47,713 candidate hospitalizations records. Note that for the very first rule, instead of performing exact matching for all variables, we

Table 3. *Uniqueness level of hospitalizations and registry databases based on different linkage rules.*

Rule	Hospital	Gender	Age	Discharge Date	Admission Date	Hospitalizations Uniqueness	Registry Uniqueness
1*	Exactly Agree	Exactly Agree	Disagree by at most 1 year*	Exactly Agree	Exactly Agree	99.74%	99.32%
2	Exactly Agree	Exactly Agree	Disagree by at most 1 year	Disagree by at most 1 day	Exactly Agree	99.21%	98.86%
3	Exactly Agree	Exactly Agree	Disagree by at most 1 year	Exactly Agree	Disagree by at most 1 day	99.23%	98.79%
4	Exactly Agree	Exactly Agree	Disagree by at most 1 year	Disagree by at most 1 day	Disagree by at most 1 day	97.82%	97.60%

*A record is considered unique for this rule if no other record matches its hospital name, admission date, patient gender, discharge date and age, even though two ages differing by one year are considered as the same age. When regarded as a linkage rule, the rule means that two records from different databases are referring to the same hospitalization if they have the same hospital name, admission date, patient gender, discharge date and patient age, while two ages differing by one year are considered as the same age.

allowed 1 year flexibility in the age variable to incorporate uncertainty resulting from different age calculation methods in hospitalizations and registry databases. As a sensitivity analysis, we also examined the linkage process with an additional rule with all variables matched exactly to be applied at the very beginning of the linkage, and the resulting links are almost identical.

In the absence of a gold-standard record identifier across the source databases, we evaluate the linkage quality by calculating the proportions of agreement among linked records on common data fields not used for linkage. For our hospitalizations and registry databases, such validation variables include the primary diagnosis code (ICD-9-CM), patient disposition, and tissue plasminogen activator (tPA) administration status. As these variables can take on missing values, we defined the proportion of agreement as the number of the linked pairs with agreed values of the validation variables over the number of the linked pairs with non-missing values of the validation variables on both source databases, assuming that values are missing independent of the linkage status. To increase the discriminative power of the validation, we also evaluated agreement on the three validating variables simultaneously by constructing a composite validator. The detailed validation results are shown in the Supplemental Table S1. Agreement rates on the diagnosis code, patient disposition, tPA status and composite validator were 96.9%, 95.4%, 98.9% and 91.5% respectively, consistent with a good linkage performance. We also evaluated statistical significance of our observed agreement by examining the agreement distribution from randomly linked hospitalizations-registry records. We simulated 5000 randomly linked datasets and calculated their agreement rate for each validator. The simulated distributions are shown in Supplemental Figure S1. Our observed agreement exceeds the maximum value obtained among random replicates for all validators (i.e., p-values < 0.001, shown in Supplemental Table S1), demonstrating that our observed agreement rates are significantly different from what would be expected by chance.

To evaluate the impact of measurement bias in the linked data related to unmatched or mismatched records, we have also compared the demographics and several key clinical variables for matched versus unmatched records with respect to the registry database. We assessed the balance in variables with standardized mean

Table 4. *Hospitalizations-registry linkage by rules.*

Rule	*Registry Uniqueness**	*Hospitalizations Records to be linked*	*Registry Records to be linked*	*Linked Record Pairs*
1	99.32%	47,713	43,487	36,724
2	98.86%	10,989	6,763	325
3	98.79%	10,664	6,438	1,413
4	97.60%	9,251	5,025	31
Total				38,493

*In the table, we demonstrate ranking the rules using registry uniqueness level. Rules were also ranked by hospitalizations uniqueness for sensitivity checking, and the resulting linked data were similar.

differences (SMDs). We found that most variables are balanced (SMD less than 0.2), and among the few unbalanced variables (age and race/ethnicity) the SMDs are all less than 0.3 (Supplemental Table S2).

Hospitalizations-vital linkage

Identifiers that are common in both hospitalizations and vital databases are patient gender, patient date of birth and zip code of patient permanent address. However, before assessing the feasibility of linkage with these identifiers, we first needed to address the problem that hospitalizations and vital databases were on different record levels – encounter level in the hospitalizations database versus individual level in the vital database.

To address this, we reduced the hospitalizations data into an individual-level database, taking advantage of the unique identification number attached to each hospitalizations record. Specifically, we consolidated multiple records of each identified individual into a single record using a long-to-wide transformation. When multiple different values existed for a single patient on identifiers that would typically not change over time (e.g., gender and date of birth), we stored all of these values in additional columns of the reduced hospitalizations database, as they imply potential data mismeasurements and are likely to be useful in the linkage. For variables whose values may change over time (e.g., zip code and discharge date), we retained the latest values of the variables to be matched to vital records. The hospitalizations database was reduced from 69,282 hospitalizations to 60,805 individuals, among which 34 individuals have 2 different date of birth values (though none with more than 2 different date of birth values), and no one has more than 1 gender value.

After the hospitalizations data were transformed to the patient level, we assessed the feasibility of the hospitalizations-vital linkage. Combinations of gender, date of birth, and zip code on their distinct values uniquely identified 99.1% of the records in the reduced hospitalizations database and 95.7% of the records in the vital database, supporting feasibility of the linkage. We tested linkage rules that allowed for flexibility in values (similar to the flexibility rules used in the hospitalization-registry linkage and outlined in Table 3). However, combinations of the identifiers allowing flexibility in their values all resulted in uniqueness less than 95%, and thus were not used as additional rules for implementing the linkage.

We then linked the reduced hospitalizations database with the vital database on distinct values of gender, date of birth, and zip code, while adding an additional constraint that the discharge date of a reduced hospitalizations records should be no later than the death date of a vital record. While searching vital records to link to each reduced hospitalizations record, we gave the highest priority to the reduced hospitalizations records with a unique date of birth value. Vital records left unlinked then became available to pair with reduced hospitalizations records with multiple values of date of birth, among which one value was randomly picked to participate in the linkage. Additional values of date of birth were used successively to link to remaining vital records if the corresponding reduced hospitalizations record had not linked in previous steps. This linkage scheme is illustrated in Figure 2. In total, we identified 16,610 ischemic stroke patients who were discharged between 10/1/2007 and 9/30/2015 and died between 1/1/2010 and 3/6/2017; this corresponds to 27.3% of the 60,805 individuals in the reduced hospitalizations data.

As a result of the hospitalizations-vital linkage, each linked vital record acquired a unique patient identification number from the hospitalizations data. The 16,610 vital records were then joined with the original encounter-level hospitalizations database on their patient identification number to transform the linked file from individual-level back to encounter-level, because one patient can have multiple hospitalizations records. This join completed the hospitalizations-vital linkage, making 19,290 (27.8%) of the hospitalizations records linked to the vital database.

We evaluated the hospitalizations-vital linkage by calculating the percentage of deaths identified in the linkage among the hospitalizations records with a discharge disposition of death. We used patient discharge disposition from the hospitalizations data as the gold standard to evaluate the linkage quality.

Among 2,776 hospitalizations records with a discharge disposition of death after 1/1/2010, 2,045 (73.7%) were linked to unique vital records. Among the 2,045 linked records, most (2,031, 99.3%) have their discharge dates and death date consistent within one day.

The evaluation for the quality of hospitalizations-vital linkage was also performed through a sensitivity checking where we reproduced the entire hospitalization-vital linkage except that the restriction that the discharge date has to be no later than the date of death was removed during the linkage but then used during the validation. Demonstrated in Supplemental Figure S2, this alternative linkage process resulted in the total of 16,529 individual-level links that were then transformed to 19,171 encounter-level hospitalizations-vital links, among which 97.3% of the links have their discharge date earlier than or equal to the date of death. We also compared the demographics and several key clinical variables for matched versus unmatched records with respect to the administrative database among those who died in hospitals to evaluate the impact of potential unmatched and mismatched records. Again, we assessed the balance in variables with SMDs and found that all variables are balanced (Supplemental Table S3).

Integration of hospitalizations, registry and vital databases

An integrated hospitalizations-registry-vital database would cover all the data fields from the three source

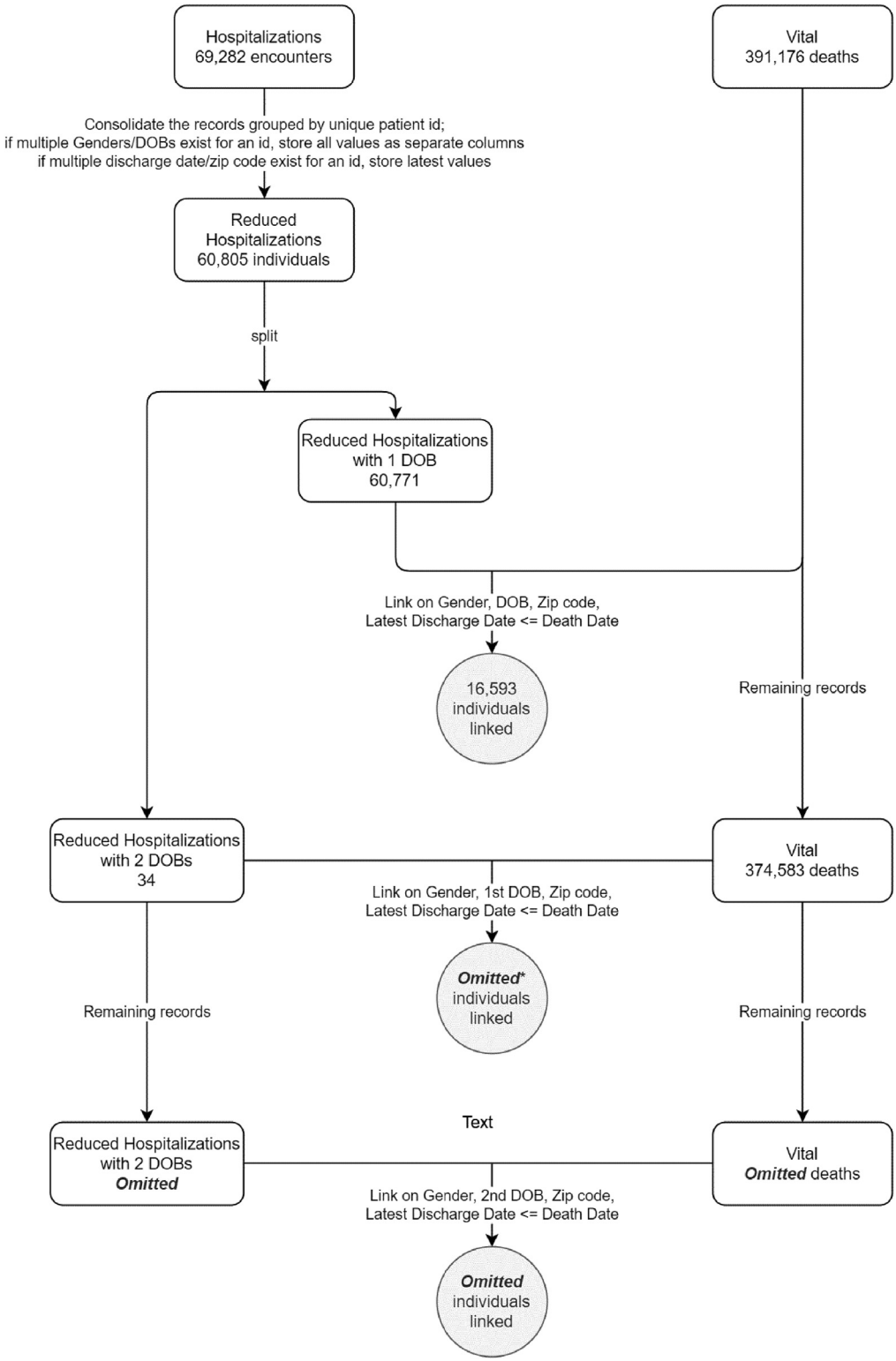

Fig. 2. *Hospitalizations-vital linkage process. Legend: *Small counts (< 11) and numbers directly calculated from such counts were omitted to comply with the local institutional review board criteria to protect patient identities.*

databases for their intersected population, i.e., all ischemic stroke hospitalizations of patients discharged between 1/1/2008 and 9/30/2015 from the 49 registry-participating hospitals. Therefore, the integration is equivalent to left-joining the linked hospitalizations-registry to the linked hospitalizations-vital data, both of which are on encounter-level. Records in the integrated database correspond to the 38,493 hospitalizations linked to registry records,

Table 5. *Reference steps for integrating regional hospitalizations, registry and vital data.*

Step 1 Integration Preparation
1) Identify the master data file
2) Determine 3 integration phases: two linkage phases involving the master file followed by the integration of the three databases
3) Determine data population for each phase
Step 2 Individual Linkage Phases (implement twice for linkages of choices)
1) If an individual-level database is involved, reduce the other (encounter-level) database into an individual-level database
2) Create linkage rules based on indirect identifiers
3) Check feasibility for each rule by calculating their corresponding data uniqueness
4) If feasible, perform rule-based linkages successively in descending order of uniqueness
5) if the linkage is on individual-level, link on extra values of static individual-level identifiers (e.g. age, gender, etc.), if any
6) If the linkage was performed on individual level, transform the linked data back to encounter level
7) Evaluate the linkage quality using additional variable(s) common to source databases or a gold standard database crosswalk, if available
Step 3 Integration of the three databases
1) Left join the linked data without the vital component to the linked data with the vital component, if the linkage qualities are acceptable in the last two phases

among which 10,660 (27.7%) were linked to vital records. The detailed procedures of our entire integration process are summarized in Table 5 and may serve as a reference to replicate hospitalizations-registry-vital integrations in alternative contexts.

Discussion

We described in detail the procedures used for integration of three distinct databases in the absence of direct record identifiers for the development of a statewide stroke database. The final integration database is composed of hospitalizations, registry and vital data over multiple years, corresponding to over 80% of statewide stroke encounters identified in the comprehensive hospitalizations data. Our integrated database, in particular, may be useful in a variety of state-level policy and surveillance investigations, including but not limited to:

1) the use of quasi-experimental design strategies in the evaluation of state-wide quality improvement initiatives.
2) the comparison of mortality or other hospital characteristics among Coverdell-participating versus non-Coverdell hospitals at the state level.
3) the modeling of outcomes among the Coverdell-participating hospitals adjusting for full range of comorbidities and procedures over time.
4) examining the impact of state policy changes (e.g., regionalization or prehospital routing policy) on patient-centered outcomes (e.g., long-term disability, mortality).

The integration process represents an invaluable mechanism to study longer-term patient outcome and mortality by harnessing the important clinical details included in registry data alongside the longer-term outcomes accessible in administrative records.

Previous reports have described hospitalizations-registry and hospitalizations-vital linkages separately with indirect identifiers. Our study adds to the literature by introducing a framework of integrating the three distinct types of data into a comprehensive database system encompassing individuals that would constitute a complete cohort based on their primary diagnosis. In addition to its three-way linkage nature, our framework distinguishes from other data linkage practices documented in existing literature in the following three aspects.

First, we emphasize determination of the target populations of each integration phases prior to implementation, which helps the investigators to focus on the maximum number of linked records that can be achieved during that particular phase, regardless of whether the linked data includes information from all three data sources. We chose not to truncate data to the smallest overlapping group of dates, because there is still potential that the earlier hospitalization data may be useful to answer questions only requiring hospitalizations and vital information. By maintaining encounters that occurred outside of the fully overlapping time period, we created a dataset to efficiently exploit the data available at any given time. However, If the investigators focus solely on research contexts that require full information from all three data sources, then one may simply focus on the fully overlapping data.

Secondly, we underscore the value in formalizing the evaluation of linkage feasibility by calculating data uniqueness in individual datasets prior to linkage. A linkage is not feasible if the indirect identifiers under consideration do not contain enough information to distinguish records in either source database, as this would lead to uncertainty on which records truly identify a particular encounter or patient. While several measurements exist to estimate the information contained in the identifiers,[11] we regard data uniqueness as the most intuitive measurement, because it directly demonstrates the percentages of records that can be distinguished in a database. While the uniqueness calculation was introduced in existing data linkage literature,[3] it was not formalized as a feasibility checking step. In our study, we used 95% uniqueness value as the threshold to assess linkage feasibility. We note that this threshold is somewhat arbitrary and subject to sensitivity checking.

Thirdly, it is important to have a process for reconciling the level of data when source database are of different natures (i.e., encounter-level versus patient-level). Specifically, our linkage process required two data transformations during the hospitalizations-vital linkage phase. The first transformation reduced the hospitalizations data into an individual-level database so that it could be linked to the vital database on individual-level identifiers. The second transformation reverted the patient population into a hospitalizations population so that the hospitalizations-vital data could be joined with the linked hospitalizations-registry data on the encounter level. Documenting the detailed process of these transformation enables tracking of the varying population throughout the entire integration process.

We also present several additional important considerations during an integration process. First is the identification of a universal master database at the onset of the integration effort. This is essential for setting up reasonable integration phases and the phase-specific data populations. Secondly, the method to address non-one-to-one links during linkage phases has to be specified. As in our study, records that are matched but not distinguishable from one another given the combination of specific identifier values will lead to non-one-to-one links. For example, several hospitalizations records are linked to one registry record when the combination of corresponding identifier values is unique in registry but not hospitalizations database. These non-one-to-one links are discarded. In the presence of high data uniqueness, this approach would not result in too much data loss. Alternatively, an investigator may also check non-one-to-one links associated with the same set of identifier values and identify the best link among all possible links through clerical review. However, the specific criteria used in manual checking may be subjective and vary across data contexts.

There are several reasons that we were unable to link all registry records with all-payer hospitalizations records for the same data collection window and facilities. First, both candidate hospitalizations and registry records to be linked were the subset of records with a primary diagnosis code associated with ischemic stroke, and the codes may be reported separately in hospitalizations and registry database for the same hospitalization. Thus, an ischemic stroke diagnosis code for a hospitalization may appear as the primary diagnosis in registry data but not the hospitalizations data. A second potential reason is that registry data are often entered manually by trained chart abstractors, and while generally of very high quality,[12] the registry may contain data entry errors contributing to unmatched identifier values. In contrast, for the hospitalizations-vital linkage, it would be unreasonable to expect each hospitalizations record to be linked to a vital record given that many stroke patients remain alive at the time of linkage, and one would not know how many hospitalizations records should be linked without a gold standard vital status variable attached to the hospitalizations database.

Our study has limitations. For the hospitalizations-registry linkage, we detected approximately 20% of hospitalizations records that were not linked to registry data. Given that the Coverdell registry is rigorously maintained, this may be indicative of inherent misclassification based on diagnosis codes in the hospitalization database. Further study is needed for validation and adjustment in using the diagnosis codes for defining stroke admissions in the local hospitalizations data. For hospitalization-vital linkage, without knowing the true vital status for all the hospitalizations records, it is hard to comprehensively validate the hospitalizations-vital links. Furthermore, we could only capture deaths that occurred after 1/1/2010 due to data availability. While we performed the linkage quality evaluation on the subset of the patients with in-hospital death after 1/1/2010, this can only serve as a proxied but not perfect evaluation for the entire hospitalizations data. For both the hospitalizations-vital and hospitalizations-registry linkages, data integration with indirect identifiers may also lead to a certain number of incorrectly matched links, which in turn may result in measurement bias through misclassification in studies using the integrated data. Reassuringly, we found relative balance between matched and unmatched records in terms of patient demographics and several key clinical variables. Nonetheless, we recommend users of the integrated database reproducing this balance checking for all variables of interest particular to their study aim to explicitly address any potential measurement bias in any given analysis. In addition, in removing non-one-to-one links, the integration may result in some missed true links (e.g., undercounting death event), even though the number of missed links should be small given the relatively high uniqueness level of our matching variable combinations (i.e., $> 97\%$ for the hospitalizations-registry linkage and $> 95\%$ for the hospitalization-vital linkage).

Conclusion

A high-quality integration of the regional hospitalizations, registry and vital data specific to ischemic stroke patients is achievable based on data linkages with indirect identifiers. Beyond our specific data results, the process and framework of the hospitalizations-registry-vital integration can be referred in other geographical and clinical context, facilitating health service researchers to answer a wide range of questions regarding long-term patient outcomes.

Funding

Agency for Healthcare Research and Quality (K08 HS024561, Zachrison) and Centers for Disease Control and Prevention (CDC) under the Paul Coverdell National Acute Stroke Prevention Program (NU58DP006072).

Declaration of Competing Interest

ZY has no conflicts to report. VN, GS and AN report employment by the Massachusetts Department of Public Health. LHS reports relationships relevant to research grants or companies that manufacture thrombolysis or thrombectomy products even if the interaction involves non-thrombolysis products: scientific consultant regarding trial design and conduct to Genentech (late-window thrombolysis) and steering committee membership (TIMELESS NCT03785678); consultant to LifeImage and Massachusetts Department of Public Health; member of Data Safety Monitoring Boards (DSMB) for Penumbra (MIND NCT03342664) and Diffusion Pharma PHAST-TSC NCT03763929); National PI for Medtronic (Stroke AF NCT02700945); National Co-PI, late window thrombolysis trial, NINDS (P50NS051343, MR WITNESS NCT01282242; alteplase provided free of charge to Massachusetts General Hospital and supplemental per-patient payments to participating sites by Genentech); Site PI, StrokeNet Network NINDS (New England Regional Coordinating Center U24NS107243). KSZ reports grants from the Agency for Healthcare Research and Quality during the conduct of the study; grants from NIH/NINDS, other from the American College of Emergency Physicians, and grants from CRICO outside the submitted work.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jstrokecerebrovasdis.2021.106236.

References

1. Bustamante A, García-Berrocoso T, Rodriguez N, et al. Ischemic stroke outcome: A review of the influence of post-stroke complications within the different scenarios of stroke care. Eur J Intern Med 2016;29:9-21. https://doi.org/10.1016/j.ejim.2015.11.030.
2. Johnston SC, Mendis S, Mathers CD. Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling. Lancet Neurol 2009;8(4):345-354. https://doi.org/10.1016/S1474-4422(09)70023-7.
3. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J 2009;157(6):995-1000. https://doi.org/10.1016/j.ahj.2009.04.002.
4. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 2002;40(8 Suppl). https://doi.org/10.1097/01.MLR.0000020942.47004.03.
5. Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Serv Res 2006;6(1):1-10. https://doi.org/10.1186/1472-6963-6-48.
6. Herrchen B, Gould JB, Nesbitt TS. Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. Comput Biomed Res 1997;30(4):290-305. https://doi.org/10.1006/cbmr.1997.1448.
7. Center for Health Information and Analysis (CHIA). The Massachusetts Acute Hospital Case Mix Database. CHIA: https://www.chiamass.gov/chia-data/. Accessed June 15, 2021.
8. Wattigney WA, Croft JB, Mensah GA, et al. Establishing data elements for the Paul Coverdell National Acute Stroke Registry. Part 1: Proceedings of an expert panel. Stroke 2003;34(1):151-156. https://doi.org/10.1161/01.STR.0000048160.41821.B5.
9. Commonwealth of Massachusetts Department of Public Health. Registry of Vital Records and Statistics. https://www.mass.gov/orgs/registry-of-vital-records-and-statistics. Accessed June 15, 2021.
10. Newcombe HB. Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford University Press, Inc.; 1988.
11. Roos LL, Wajda A. Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods Inf Med 1991;30(2):117-123. https://doi.org/10.1055/s-0038-1634828.
12. Xian Y, Fonarow GC, Reeves MJ, et al. Data quality in the American Heart Association Get with the Guidelines-Stroke (GWTG-Stroke): Results from a National Data Validation Audit. Am Heart J 2012;163(3). https://doi.org/10.1016/j.ahj.2011.12.012.