Article published online: 2021-02-10

Validating the Matching of Patients in the Linkage of a Large Hospital System's EHR with State and National Death Databases

Rebecca B. N. Conway[1,*] Matthew G. Armistead[2] Michael J. Denney[2] Gordon S. Smith[3]

[1] Department of Community Health, University of Texas Health Science Center at Tyler, Tyler, Texas, United States
[2] Department of Biomedical Informatics, West Virginia Clinical and Translational Science Institute, Morgantown, West Virginia, United States
[3] Department of Epidemiology, West Virginia University, Morgantown, West Virginia, United States

Address for correspondence Rebecca Baqiyyah Conway, PhD, American Academy of Epidemiology, 2008 S. Wiley Avenue, Tyler, Texas 75701, United States (e-mail: rebeccabnconway@yahoo.com).

Appl Clin Inform 2021;12:82–89.

Abstract

Background Though electronic health record (EHR) data have been linked to national and state death registries, such linkages have rarely been validated for an entire hospital system's EHR.

Objectives The aim of the study is to validate West Virginia University Medicine's (WVU Medicine) linkage of its EHR to three external death registries: the Social Security Death Masterfile (SSDMF), the national death index (NDI), the West Virginia Department of Health and Human Resources (DHHR).

Methods Probabilistic matching was used to link patients to NDI and deterministic matching for the SSDMF and DHHR vital statistics records (WVDMF). In subanalysis, we used deaths recorded in Epic ($n = 30{,}217$) to further validate a subset of deaths captured by the SSDMF, NDI, and WVDMF.

Results Of the deaths captured by the SSDMF, 59.8 and 68.5% were captured by NDI and WVDMF, respectively; for deaths captured by NDI this co-capture rate was 80 and 78%, respectively, for the SSDMF and WVDMF. Kappa statistics were strongest for NDI and WVDMF (61.2%) and NDI and SSDMF (60.6%) and weakest for SSDMF and WVDMF (27.9%). Of deaths recorded in Epic, 84.3, 85.5, and 84.4% were captured by SSDMF, NDI, and WVDMF, respectively. Less than 2% of patients' deaths recorded in Epic were not found in any of the death registries. Finally, approximately 0.2% of "decedents" in any death registry re-emerged in Epic at least 6 months after their death date, a very small percentage and thus further validating the linkages.

Conclusion NDI had greatest validity in capturing deaths in our EHR. As a similar, though slightly less capture and agreement rate in identifying deaths is observed for SSDMF and state vital statistics records, these registries may be reasonable alternatives to NDI for research and quality assurance studies utilizing entire EHRs from large hospital systems. Investigators should also be aware that there will be a very tiny fraction of "dead" patients re-emerging in the EHR.

Keywords
- SSDMF
- NDI
- vital statistics
- mortality
- EHR
- validation
- linkage
- social security number
- data analysis

[*] Former affiliation: University of Texas Health Science Center at Tyler, School of Community and Rural Health, Tyler, Texas, United States.

received
July 2, 2020
accepted after revision
November 16, 2020

DOI https://doi.org/10.1055/s-0040-1722220.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Background and Significance

With the widespread implementation of repurposed electronic health record (EHR) data for clinical, population, and quality assurance research, the use of repurposed EHR data is transforming the way research with human subjects in health sciences is conducted. Mortality is one of the most important health outcomes, and certainly the objective end point, in medical research, and EHR data are frequently linked to state or national death registries to determine outcomes.[1,2] Though these linkages are usually on an individual investigator project basis, this is also beginning to occur with entire EHRs and clinical data warehouses. The potential benefit in terms of knowledge gained in the epidemiology of many acute and chronic illnesses, comparative effectiveness studies, quality assurance, improved clinical practice, and subsequent lives saved is huge. However, the validity of such linkages has rarely been determined, particularly of entire EHR systems.

With the institution of the 2011 Social Security Administration's (SSA) ruling that it could not release state owned data to the Social Security Death Masterfile (SSDMF),[3,4] changes in the completeness of the SSDMF, and a resulting great reduction in the number of deaths included in the releasable file,[5,6] the need for understanding the validity of linkages with other major sources of death data used for research has increased. Our work helps address uncertainty about linkages with these death databases and thus strengthens clinical and health policy decisions arising from their usage. Further, accurately capturing patient mortality in quality measures[7] is especially important for follow-up after hospital discharge.

Objectives

We conducted a validation study of West Virginia University Medicine's (WVU Medicine) linkage of its entire EHR with three external sources of death data: the West Virginia State Department of Health and Human Resources (DHHR) vital statistics files (WVDMF), the SSDMF, and the national death index (NDI) for the years 1994 to 2015 and compared the agreement between these three sources and that recorded in the EHR itself. The purpose of this study was threefold: (1) to determine the validity of the linkages to these three standard external death sources; (2) to determine the differences, that is the potential gain/loss in benefit in using state vital statistics data or the SSDMF versus NDI data; and (3) to characterize patients incorrectly linked, especially those who reoccur in the EHR after being reported deceased, our so called "walking dead." We have previously validated the extract, transform, and load process used to populate WVU Medicine's clinical data warehouse.[8] The current study is a continuation of our validation processes of the data used in this warehouse.

Methods

Overview of the Death Databases

The SSDMF is a public use registry of death records that was created by the SSA to be compliant with Freedom of Information Act requests from the public.[9,10] It contains the social security numbers (SSNs), first and last names, and dates of births and deaths of individuals that the SSA knows to be deceased. Because of data ownership by individual U.S. states, inadvertent disclosure of SSNs of living individuals, and other breaches of privacy rights of nondeceased U.S. citizens, at the beginning of 2011 the SSA began putting greater restrictions on the use of the SSDMF,[3,4] resulting in an exclusion of 40% of its death reports that came from state vital statistics offices.

The NDI is a national mortality registry maintained by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). It is a centralized database of deaths reported to state vital statistics offices. The NDI data are made available for statistical health research and comparative effectiveness analysis only.[11] In contrast to the SSDMF of the Social Security Administration, linkage of NDI data with external sources of data is done at NCHS by NCHS staff. Thus, the NDI user must send their patient or study participant identifying data to NCHS.

The West Virginia DHHR death file, hereafter referred to as the West Virginia Death Masterfile (WVDMF), contains vital records of all deaths occurring in West Virginia. Mortality data are entered from death certificates and contains identifying data such as name, date of birth, sex, social security number, place of residence, next of kin, father, mother, date of death, and underlying and contributing causes of death

Epic is an electronic medical record system developed by Epic Systems Corporation (Verona, Wisconsin), a privately held health care software company. Epic is WVU Medicine's operational database, from which data are extracted for financial reporting quality assurance and research efforts. Medsite is a clinical result reporting system developed by WVU Medicine information technology staff (M.J.D. and Jeff Cox) and was Epic's predecessor system.[8]

Linkage Methods

Social Security Death Master File

In August of 2015, we obtained the SSDMF, containing 92,279,854 records and supplements for the next 2 years. To be consistent with data from NDI, only deaths up to December 31, 2015 were used. The SSDMF was linked with the WVU Medicine's EHR (Epic), which contained 1,050,627 individual records using a deterministic algorithm, detailed in ►Supplementary Appendix A (available in the online version). This algorithm had been developed for previous projects requiring linkage of WVU Medicine's EHR data (from both Epic and Medsite) to other third party datasets. A deterministic linking method, using discrete identifiers such as name and/or SSNs, is appropriate for the data environment investigated here, wherein identifying data are of good quality[12] (see ►Supplementary Appendix B, available in the online version for a detailed description of our deterministic algorithm). Key identifiers used for the linkage included SSNs, first names, last names, middle names, and birth dates. SSDMF records with complete month and year of death or birth were ascribed on the 15th of the respective month as the day of death, where the day was missing in the

death date, or as the day of birth where it was missing for the birthdate year. This was done to calculate a close estimate for age at death. For those for whom only the birth year was available, we assigned July 1st of their birth year as their date of birth. This too was done to facilitate matching and computing an estimated age at death.

We used an internal scoring method to rank the accuracy of the birth and death dates based on missingness and improbabilities in the SSDMF birth and death dates. Birth dates with a missing SSDMF of a specific day of the month but assigned the 15th day of their birth month by us, were given a birthdate score of 4. Birth dates assigned July 1st because of missing birth month and day of month were assigned a birth date score of 3. Death dates with a missing SSDMF of a specific day of the month but assigned the 15th day of their death month by us, were given a death date score of 4. Complete valid dates were given a score of 5 (This scoring system is detailed in ►**Supplementary Appendix B**, available in the online version).

When linking individuals, we used birthdates that were within 3 years of each other in the two matching sources, requisite of the first three letters of the first name being identical and the first two letters of the last name being identical, or the use of Soundex match on the first and last names (https://www.archives.gov/research/census/soundex.html) (see ►**Supplementary Appendix C**, available in the online version).

National Death Index Data

In May of 2017, 1,050,627 WVU Medicine Epic patient names and demographic identifiers, including SSN were sent to the NCHS to link with the NDI's death record database. Criteria used in the export of the individual patient's identifiers were (1) that the patient had to have had at least one reported visit between January 1, 1994 and December 31, 2015; and (2) the patient did not have a subsequent visit after December 31, 2015.

NDI uses a two-stage matching scheme to link the data. In the first step, possible matches are selected by meeting any one of the following seven criteria: (1) SSN; (2) exact month and plus/minus 1 year of birth, first and last name; (3) exact month and plus/minus 1 year of birth, first and middle initials, last name; (4) exact month and day of birth, first and last name; (5) exact month and day of birth, first and middle initials, last name; (6) exact month and year of birth, first name, father's surname; (7) (for female decedents) exact month and year of death, first name, last name, and father's surname. In the second stage of the matching process, a probabilistic score is assigned to the possible matches identified in stage 1.

West Virginia Health and Human Resources Death Data

In September of 2017, we obtained the WVDMF from the West Virginia Department of Health and Human Resources (DHHR) to link to our EHR. The WVDMF contained 482,003 records, one record for each death reported in West Virginia between 1994 and 2016. The linkage between the WVDMF and the WVU Medicine EHR was based on a deterministic matching scheme similar to that used for the SSDMF.

Epic Death Data

We used clarity generated data from Epic to determine the deaths listed in our EHR. There were 54,129 records in clarity recorded as a death, with either a death date provided, or a patient field status code listed as deceased. Twenty-five thousand of these only had a patient status field code of deceased, but no date of death listed. Less than 20 had a death date but not a field status code of deceased. Of the 54,129 patients listed as deceased, 318 did not have a social security number.

Reliability Statistics

Kappa statistics[13,14] was conducted to compare mortality capture statistics of the linkages of the three external death databases. Due to the changes in usage of the SSDMF in 2011, we ran sensitivity and specificity tests for two time points to see how this affected our results. The time periods chosen were deaths up to December 31, 2009 and then again through December 31, 2015. From each of the years 2009 and 2015, we randomly sampled 100 gold standard (GS) dead patients and 100 GS alive patients. A GS dead patient was a patient who had a hospital admission with a discharge disposition of EXPIRED, with a valid death date, and the death date was within the bounds of that hospital admission's admission and discharge date/time. A GS alive patient was a patient who had a visit and a subsequent last known visit at least 30 days after that initial visit.

Results

Of the patients matched to the SSDMF, 88% were deemed valid using our deterministic algorithm (see ►**Supplementary Appendix A**, available in the online version). The remaining 12% were clearly invalid or of questionable validity, for example, having very different first names or mismatched last names and/or birthdates or implausibly early death dates. The final number of WVU Medicine patients matched to the SSDMF as deceased was 254,929. Similar findings were seen for the WVDMF. Of the 1,050,627 patients sent to the NDI, 189,082 had a match with a status code of 1 using NDI's probabilistic matching. Of these, 178,132 (94%) had a match on SSN and/or other matching fields (NDI class code 2).

The sex and age distributions of the 1,050,627 WVU Medicine patients linked to the SSDMF, the NDI, and WVDMF are presented in ►**Table 1**. There was a generally similar distribution of males and females overall and by age group, with the exception of those 85 years and older. There were 41% more females among those 85 years and older.

The number of WVU Medicine deceased patients matched to each of the three external death registries and the number captured in the EHR, as well as the percentage co-captured by the other three databases are presented in ►**Table 2**. The SSDMF had the highest number of matches, followed by the WVDMF. WVU Medicine's EHR captured the fewest number of decedents. When using patients recorded in WVU Medicine's EHR as deceased as the reference, each of the three external death registries had a similar proportion of matches

Table 1 Sex and gender distribution of the EHR patient population sent to NDI

Overall	1,050,627
Male	517,568
< 1 y	517,568
1–21 y	145,401
22–64 y	263,327
65–84 y	263,327
85 y and older	263,327
Female	523,194
< 1 y	17,449
1–21 y	128,533
22–64 y	269,614
65–84 y	80,547
85 y and older	27,031
Unknown	9,284
< 1 y	968
1–21 y	2,820
22–64 y	2,895
65–84 y	459
85 y and older	2,142

Abbreviations: EHR, electronic health record; NDI, national death index.
Note: Eighty-eight males and 75 females had missing information on date of birth.

to their death database, with match "rates" ranging from 84.3 to 85.5%. However, when not restricted to the 30,217 patients recorded as deceased in the EHR, the percentage of co-captured decedents was not as high among the three external death databases. For example, 59.8% of the decedents captured by the SSDMF were also captured by the NDI. Conversely, 80% of the decedents captured by the NDI were also captured and matched to the SSDMF. Of the 189,095 matches to the NDI and the 208,297 matches to the WVDMF, approximately 52,000 of the matches in NDI were not in the WVDMF. However, approximately 50,000 of those deaths were deaths recorded in other states.

Kappa statistics suggested that the death registry with highest level of agreement with the other two registries was the NDI, with moderate to substantial agreement between the NDI and the other two death registries. Rates of agreement were as follows: SSDMF and NDI = 60.6%, representing moderate to substantial agreement; SSDMF and WVDMF = 27.9%, representing fair agreement; and NDI and WVDMF = 61.2%, representing substantial agreement.[13,14] The detailed calculation of our kappa statistics for these data is presented in ►Supplementary Appendix D, ►Table A3 (available in the online version).

Our sensitivity analyses and specificity analysis revealed that both time periods, i.e., prior to 2010 and 2016 and after, had a sensitivity of 99% and a specificity of 100%. For the December 31, 2009 cutoff point, there was one false negative. This patient had an office visit with a check in and check out time but was documented as deceased in the DHHR death Masterfile with a death date 281 days prior. For the December 31, 2015 cutoff point, there was one false negative patient who had a nephrology visit, including a check-in and check-out time, but two death sources (Epic and the DHHR death Masterfile) recorded this patient as having died 6 days prior.

►Table 3 shows the percentage of patients, by age and sex, recorded in the EHR as deceased but matched to only one of the three external death registries or not matched to any of the external databases. Between 2.1 and 3.0% of the deceased patients overall were only matched to one of the three external databases, though this varied slightly by age group. The percentage was highest for infants, followed by children and adolescents up to age 21, with the NDI capturing the greatest percentage unmatched to any of the two other external death registries. No differences by sex were observed. Two percent of patients recorded as deceased in the EHR were not matched to any of the external databases, with the greatest percentage occurring during the first year of life. Among these patients, there was higher prevalence of invalid social security numbers, 29.6% versus approximately 4% for the EHR overall (data not depicted). There was also a higher prevalence of apparently false names, specifically "John Doe" or "Jane Doe," 2.0 versus 1% for the EHR overall (data not depicted).

Though some patients recorded as deceased in at least one of the four death databases reappeared in the EHR at least 6 months after their recorded date of death (►Table 4), the percentage reappearing in the EHR was similar regardless of the database identifying them as deceased, with percentages ranging from 0.19% in the WVDMF to 0.29% in the NDI. For both Epic and NDI these tended to be male deaths, but there was a similar proportion by sex for the WVDMF. Median age at death in this group ranged from 63 to 70 years in the death databases. Patients matched to the WVDMF but reappearing at least 6 months after their recorded death date had the highest average number of visits recorded in the EHR.

Table 2 WVU medicine decedents from the four death data sources and amount co-captured between databases, % (*n*)

	Total deaths identified (*n*)	Deaths co-captured by the death databases, % (*n*)		
Death database		**SSDMF**	**NDI**	**WVDMF**
Epic/MedSite	30,217	84.3% (25,460)	85.5% (25,823)	84.4% (25,492)
SSDMF	254,929		59.8% (152,658)	68.5% (174,718)
NDI	189,095	80% (152,658)		77.9% (136,201)
WVDMF	208,297	83.9% (174,718)	65.4% (136,201)	

Abbreviations: NDI, national death index; SSDMF, Social Security Death Masterfile; WVDMF, West Virginia Death Masterfile.

Table 3 Patients recorded as deceased in Epic but unmatched to external death databases

	Number of patients recorded in Epic as deceased	Percentage (%) of patients recorded in Epic as deceased and matched to one but not the other two death registries			Found in neither SSDMF, NDI, nor WVDMF
		SSDMF	NDI	WVDMF	
Overall	30,217	2.7	3.0	2.1	2.1
Age group					
Men	15,933	2.5	3.3	2.0	2.0
< 1 y	234	0.4	7.7	6.4	32.5
1–21 y	313	2.2	5.3	3.2	4.8
22–64 y	5,599	2.6	3.9	2.2	2.1
65–84 y	7,453	2.7	2.9	1.7	0.9
85 y and older	2,334	1.8	2.6	1.8	1.6
Women	14,273	2.8	2.8	2.1	2.2
< 1 y	181	0.0	10.0	0.3	33.0
1–21 y	201	4.2	5.5	4.5	4.4
22–64 y	3,799	2.7	3.3	0.2	2.1
65–84 y	6,499	3.0	2.4	1.9	1.5
85 y and older	3,593	2.7	2.5	2.5	2.1
Unknown sex	11	0	0	0	36.4
< 1 y	3	0	0	0	66.7
1–21 y	0	0	0	0	0
22–64 y	1	0	0	0	100
65–84 y	5	0	0	0	20
85 y and older	2	0	0	0	0

Abbreviations: NDI, national death index; SSDMF, Social Security Death Masterfile; WVDMF, West Virginia Death Masterfile.

Table 4 Characteristics of patients appearing in the EHR at least 6 months after their reported date of death in one of the four death data sources

	Epic Total deceased $n = 30,217$	WVDMF Total deceased $n = 208,297$	NDI Total deceased $n = 189,095$
Deceased and reemerging	$n = 73$, % = 0.24	$n = 396$, % = 0.19	$n = 212$, % = 0.29
Sex, % female	42.5	50.3	39.7
Age at death, median (years)	63	70	68
Total visits	290	2,402	332
Visits per patient	3.97	6.06	1.56

Abbreviations: EHR, electronic health record; NDI, national death index; SSDMF, Social Security Death Masterfile; WVDMF, West Virginia Death Masterfile.

Discussion

We evaluated the linkage of a large hospital-based medical center's EHR with three external death registries. We found that the three death registries, the SSDMF, the NDI, and the WVDMF, identified deceased patients fairly accurately and identified them to a similar degree. However, out of the three, we found the NDI to be the most consistent with the other two death registries in classifying patients as deceased or not deceased. Finally, we observed less than one-half of 1% of patients identified as deceased by one of the three death registries or recorded as deceased within the EHR re-emerged in the EHR at least 6 months after their documented death date. We found this very small percentage of deceased and remerging patients, our so called "walking-dead," a further validation of the accuracy of our linkages. To our knowledge this is the first report to document the validation process of the linkage of an entire hospital system's EHR with state and national death registries.

The SSDMF, NDI, and WVDMF each captured approximately 85% of patients recorded as deceased in our EHR, and of these decedents the co-capture rate was approximately

85%. However, when not restricted to patients with a deceased code within the EHR, this co-capture rate among the three external databases varied considerably. While the NDI appeared to be the most stringent, capturing only 60% of deaths also captured by the SSDMF, as compared with the SSDMF's capture of 80% of deaths captured by NDI, of the three the NDI also had the highest level of agreement with the other two death registries. However, the Kappa statistic rate of agreement between NDI and the SSDMF was considerably less than the 74% agreement rate observed by Hanna et al for HIV deaths occurring in New York City between 2000 and 2004,[15] which likely reflects the change in policy of the SSDMF put in place in 2011 that restricted release of death data owned by state vital statistics departments.[3,4]

Our data also suggest that the state vital statistics data are a low-cost reasonable alternative to the NDI for states where this data can be obtained for free or at a nominal price, but is a poor replacement for the SSDMF, particularly for hospitals in cities in close proximity to multiple states, as is Morgantown, West Virginia where WVU Medicine is located. In a study assessing the validity of self- or proxy-reported family history of cancer, Rauscher and Sandler reported that state vital statistics data had much better match rates for deceased individuals than did NDI when identifying information was incomplete, especially missing data on social security number.[16] In their data, which lacked social security numbers, NDI was only able to successfully match 63% of deaths verified by state matches and only 10% not verified by state statistic records. By contrast, state vital statistic records were able to match with 95% of those matched to NDI and 55% of those not successfully matched by NDI.[16] Similarly, in the linkages to our EHR, NDI was able to successfully match 65% of deaths also matched to state vital statistics records, but in contrast to Rausher and Sandler the converse was not as similar; West Virginia state vital statistics records, i.e., the WVDMF, were only able to match 78% of those matched by NDI. Despite the modest interrater agreement between the WVDMF and the SSDMF, each of the three external death registries identified as deceased approximately 85% of patients recorded as deceased within our EHR. Sensitivity and specificity results for deaths occurring before and after the 2011 SSDMF policy change in usage suggests that the SSDMF policy change did not materially affect this.

Approximately 2% of our patient population listed as deceased within our EHR were not matched to any of the three external death registries. Zingmond et al found that hospital deaths unmatched to state death data increased with the age of the patient and tended to be female and that for patients aged 65 years and older unmatched hospital deaths were overwhelmingly female patients.[17] We did not observe this in our data. By contrast, in our population, the percentage listed as deceased but not found in any of the external death registries decreased with age up until age 85, a pattern that did not vary by sex. Deaths appearing in Epic but not linked to any of the external death databases may be due to several reasons. Still births and newborns who died before ever leaving the hospital may never have received an SSN and thus may never have been linked to the SSDMF. Zingmond et al reported that of the approximately 2,800,000 patient discharges in the state of California for a given calendar year, approximately 800,000 did not have valid SSNs. The majority of these were due to missing SSNs from neonates discharged after birth.[17] Newman and Brown, in a mortality linkage validation study of a California hospital (UCSF) linked with the California state death registry and the SSDMF found that 22% of patients who died in the hospital did not have an SSN.[18] They also noted that approximately two-thirds of hospital deaths in their institution that were unable to be matched to state and national data were among infants under 1 year of age.[18] We, too, observed that a high number of our unmatched deaths in our EHR were those of infants under 1 year of age. Many of the still births and neonates among those infants may never have received a death certificate, further explaining why they were not linked to WVDMF or NDI.

Another source of the unlinked death data recorded in Epic may be deaths of U.S. citizens who died abroad and thus the death record would be held by the U.S. Department of State and not captured by NDI or state vital statistics departments, or of non-U.S. citizens or U.S. adult citizens never obtaining SSNs. Foreign visitors or non-U.S. citizens in the country without visas may have died in the hospital but these patients may not have received a U.S. death certificate nor had SSN and thus also would not be linked to the external databases. Additional sources are elderly women who never worked outside the home as well U.S. patients who never worked legally, i.e., patients who were U.S. citizens who always worked under the grid and thus never had a social security number. A tiny minority of patients may have also provided pseudonyms[19] or made up social security numbers because either they did not want or feel the need for such personal information to be held by the health care system or because the patient did not have an SSN. In fact, approximately 2% of the unlinked deceased patients had names of either "John Doe" or "Jane Doe" and approximately 4% had obviously invalid SSNs, double the percentages of 2 and 1%, respectively, for all deaths recorded in the EHR. Lack of valid SSNs in patient records has been reported elsewhere.[17]

Our report on the linkage of state and national death registries with a large hospital in Morgantown, West Virginia, an area whose catchment population includes residents from the bordering states of Pennsylvania, Maryland, and Ohio may have relevance to other U.S. states or even different countries whose residents can fairly freely move between borders and receive health care outside of their state or country of residence, such as in the Schengen Area of the European Union where many citizens experience medical emergencies outside of their country of residence. The WVDMF only captured deaths of individuals dying in the state of West Virginia, regardless of the decedent's state of residence. Decedent information of West Virginia residents dying outside the state of West Virginia was dependent on courtesy notification by the decedent's state of death. This is captured in the lower capture statistics for the WVDMF compared with the NDI which receives death data from all 50 U.S. states and territories. Thus, hospital quality assurance and other research seeking to capture information on

hospital or post-discharge mortality should bear this in mind when a substantial portion of the patient population resides outside of the state or country of the hospital's locale. As pointed out by Brand et al[20] and Ucinski et al,[21] the importance of cross-border cooperation in medical activities cannot be understated. Accurately capturing patient mortality is especially important for follow-up after discharge, guiding important clinical and health policy decisions.

A limitation of our study was our use of different matching schemes for the NDI than for the SSDMF and the WVDMF. We were not able to procure the weights used in NDI matching algorithm, thus we could not replicate the same probabilistic matching scheme for SSDMF and WVDMF. Conversely, we were not able to apply our deterministic matching algorithm to the NDI because the NDI record linkage is done at the NCHS and they use a probabilistic matching scheme. Though other linkage approaches exist, including entity resolution, we believe our use of deterministic matching for the SSDMF and the WVDMF was appropriate given the data rich environment of our EHR.[12] Limitations of our study also include potential period effects in the validity of the linkages as the proportion of the U.S. population with SSNs increased over time. Another limitation of our study is the close proximity of WVU Medicine to the bordering states of Pennsylvania, Ohio, and Maryland. Approximately one-fourth of our patient population is from one of these three neighboring states and likely accounted for the poor inter-rater classification agreement between the SSDMF and the West Virginia state vital statistics records. Finally, scheduled visits may have made the patient appear to be in the EHR after the date of death, but the visit never occurred. We tried to account for this by allowing a lag period of up to 6 months after the recorded date of death but there may have been visits scheduled further in advance than this.

Conclusion

While commercial sources of death data exist, the most common sources for investigators and institutions remain state departments of health vital statistics records, the NDI, and the SSDMF. Our data suggest that even for large scale health care delivery quality assurance evaluations or epidemiological research with EHR data, the NDI and state vital statistics records provide reasonable levels of accuracy for large scale automated research. Only 2 to 3% of patients recorded as deceased within the EHR were not found in at least one of the external death registries and 2% were found in none of the external registries. Further, only 0.2 to 0.3% of the patients reported as deceased in one of the death registries reappeared for a patient visit at least 6 months after their recorded death date. We find this to be a low number and within the range of vicissitudes of human behavior among patients or data entry error.

Although the findings of this manuscript are exclusively conducted in U.S., it is very relevant to international audiences. In the Schengen Area, for example, countries have similar national and regional death registries. These data can certainly be linked to an institution's EHR data to cross-check the validity of vital statistics from different data sources. In fact, we expect that the accordance between similar databases from Nordic countries might be better. Those registries may also be reasonable alternatives to country-specific datasets such as the NDI for research and policy-making purposes. Similar registry data can be found in other countries or regions, such as Japan, Singapore, and Hong Kong. Although our findings could not be generalized directly to any international environments, the methodology and the findings of our study may serve as a useful reference for international researchers and quality assurance investigators.

Clinical Relevance Statement

Linking of hospital health care delivery with mortality is an acceptable hard outcome for assessing quality of care. And because of the trend toward shorter hospital stays and the transferring of sicker patients to other care institutions, overall mortality is considered a better end point than within hospital mortality.[13] The trend in the usage of repurposed health care data for research other than quality assurance has also spurred interest in linking such data to mortality registries. Assessing the validity of the linkages of EHRs and clinical data warehouses with external mortality registries may have an important impact on future health care delivery decisions and perceived knowledge, i.e., what we think we know, about the epidemiology of many acute and chronic diseases. This is also a primary step in the needed configuration of "smarter" EHRs for clinical care and population health,[22] quality assurance, and performance measurements.[23]

Multiple Choice Questions

1. What percentage of patients documented as deceased might subsequently return for care?
   a. 10%
   b. 5%
   c. <1%
   d. 20%

   Correct Answer: The correct answer is option c, <1%. The Conway et al study "observed less than one-half of 1% of patients identified as deceased by one of the three death registries (NDI, SSDMF, state death registry) or recorded as deceased within the EHR re-emerged in the EHR at least 6 months after their documented death date."

2. Of the three primary sources of death data in the United States, which is the most valid for comprehensively capturing deaths of patients in a hospital system?
   a. SSDMF
   b. NDI
   c. State vital statistics records
   d. Epic

   Correct Answer: The correct answer is option b, NDI. The Conway et al, study concludes that the NDI is "the most consistent … in classifying patients as deceased or not deceased" when compared with the SSDMF and the state death registry.

3. Of the examples of identifiers needed for data linkage, which one is incorrect?
   a. First name
   b. Date of birth
   c. Mother's last name
   d. Father's last name

Correct Answer: The correct answer is option c, mother's last name. For the NDI matching algorithm, two of the seven matching criteria sets include father's last name while (patient's) first name and date of birth are used as criteria in all three death registries (NDI,SSDMF, and the state death registry). Mother's last name is not used as a matching criterion in any of these death registries.

Protection of Human and Animal Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed by WVU University Institutional Review Board.

Funding

This study was funded by National Institute of General Medical Sciences, grants: 1U54GM104942-02, 2U54GM 104942-02 and National Institute on Drug Abuse, grants: 1UG3DA044825, R21DA040187.

Conflict of Interest

None declared.

Acknowledgments

R.B.N.C. conceived and directed the study, researched the data, and wrote the manuscript. M.G.A. and M.J.D. conceived the study and researched the data and contributed to the discussion. G.S.S. contributed to the discussion. The authors would like to thank Dr. David Zhemming Fu, Dr. Trevor Orchard, and Dr. Hillary Keenan for their helpful feedback on this manuscript.

References

1 Ballard DJ, Ogola G, Fleming NS, et al. Impact of a standardized heart failure order set on mortality, readmission, and quality and costs of care. Int J Qual Health Care 2010;22(06):437–444
2 Fleming NS, Ogola G, Ballard DJ. Implementing a standardized order set for community-acquired pneumonia: impact on mortality and cost. Jt Comm J Qual Patient Saf 2009;35(08): 414–421
3 National Technical Information Service. Important Notice: Change in Public Death Master File Records. Alexandria, VA: National Technical Service; 2011
4 da Graca B, Filardo G, Nicewander D. Consequences for healthcare quality and research of the exclusion of records from the Death Master File. Circ Cardiovasc Qual Outcomes 2013;6(01):124–128
5 Maynard C. Changes in the completeness of the social security death masterfile: a case study. The International Journey of Epidemiology. 2013;11(02). Available at: http://ispub.com/IJE/11/2/1604
6 Navar AM, Peterson ED, Steen DL, et al. Evaluation of mortality data from the social security administration death master file for clinical research. JAMA Cardiol 2019;4(04):375–379
7 D'Amore JD, Li C, McCrary L, et al. Using clinical data standards to measure quality: a new approach. Appl Clin Inform 2018;9(02): 422–431
8 Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform 2016;94:271–274
9 The Freedom of Information Act, 5 U.S.C., section 552. Department of Justice, Office of Information Policy; 1980
10 Perholtz v. Ross, Civ. No. 78–2385 and 78–2386 (U.S. District Court for the District of Columbia. 1980)
11 National Center for Health Statistics. National Death Index User's Guide. Hyattsville, MD: National Center for Health Statistics; 2013
12 Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. AHRQ Methods for effective health care. In: Linking Data for Health Services Research: A Framework and Instructional Guide. Rockville, MD: Agency for Healthcare Research and Quality (U.S.); 2014
13 Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46
14 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(03):276–282
15 Hanna DB, Pfeiffer MR, Sackoff JE, Selik RM, Begier EM, Torian LV. Comparing the national death index and the social security administration's death master file to ascertain death in HIV surveillance. Public Health Rep 2009;124(06):850–860
16 Rauscher GH, Sandler DP. Validating cancer histories in deceased relatives. Epidemiology 2005;16(02):262–265
17 Zingmond DS, Ye Z, Ettner SL, Liu H. Linking hospital discharge and death records—accuracy and sources of bias. J Clin Epidemiol 2004;57(01):21–29
18 Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. J Am Med Inform Assoc 1997;4(03):233–237
19 Rendleman N. False names. West J Med 1998;169(05):318–321
20 Brand H, Hollederer A, Wolf U, Brand A. Cross-border health activities in the Euregios: good practice for better health. Health Policy 2008;86(2-3):245–254
21 Ucinski T, Dolata G, Hełminiak R, et al. Integrating cross-border emergency medicine systems: securing future preclinical medical workforce for remote medical services. Best Pract Res Clin Anaesthesiol 2018;32(01):39–46
22 Willett DL, Kannan V, Chu L, et al. SNOMED CT concept hierarchies for sharing definitions of clinical conditions using electronic health record data. Appl Clin Inform 2018;9(03):667–682
23 Deakyne Davies SJ, Grundmeier RW, Campos DA, et al;Pediatric Emergency Care Applied Research Network. The pediatric emergency care applied research network registry: a multicenter electronic health record registry of pediatric emergency care. Appl Clin Inform 2018;9(02):366–376