5-30-2020

Extracting and Standardizing Medical Examiner Data to Improve Health

Shamsi Daneshvari Berry
Department of Health Informatics and Information Management, University of Mississippi Medical Center, Jackson, MS, USA

Heather J H Edgar
Department of Anthropology, University of New Mexico, Albuquerque, NM, USA; Office of the Medical Investigator, University of New Mexico, Albuquerque, NM, USA

Recommended Citation
Berry SD, Edgar HJH. Extracting and Standardizing Medical Examiner Data to Improve Health. AMIA Jt Summits Transl Sci Proc. 2020 May 30;2020:63-70. PMID: 32477624; PMCID: PMC7233086.

Extracting and Standardizing Medical Examiner Data to Improve Health

Shamsi Daneshvari Berry [a] and Heather J.H. Edgar [b,c]

[a] Department of Health Informatics and Information Management, University of Mississippi Medical Center, Jackson, MS, USA
[b] Department of Anthropology, University of New Mexico, Albuquerque, NM, USA
[c] Office of the Medical Investigator, University of New Mexico, Albuquerque, NM, USA

Abstract

Data from medical examiner offices are not commonly used in informatics but may contain information not in medical records. However, the vast majority of data is not standardized and is available only in large free text fields. We sought to extract information from the medical examiner database using Canary, a natural language processing tool. The text was then standardized to fit the selected normative answer list for each field. Multiple terminology and vocabulary standards from a variety of settings were utilized as data came from the medical examiner and interviews with next of kin. Thirty-seven percent of the metadata fields could be mapped directly to existing standards, twenty-five percent required a modification, and thirty-eight required creation of a standardized normative answer list. The newly formed database (New Mexico Decedent Image Database (NMDID)), will be available to researchers and educators at the beginning of 2020.

Introduction

Medical examiners investigate roughly 20% of the deaths in the United States.[1] They conduct investigations into the cause and manner of death as well as circumstances of death. Investigations into deaths are based on state laws, not federal, so the types of cases that are investigated differ on a state-to-state basis. Depending on the location, the deaths investigated include roughly one-half to two-thirds natural causes.[2] In the state of New Mexico in 2010, the majority of cases were from natural (25%), or accidental (35%) causes, with 17% suicides, 13.5% unknown causes and 9.5% homicides.[3] These data are extremely important for research on public health issues and epidemiology.[2,4-5] Additionally, medical examiner offices utilize electronic records, but they are often unstandardized and contain many free text fields. However, there is growing recognition of the need for informatics training in Pathology.[6]

Most commonly, informatics is focused on using data from the electronic medical record to improve the health of the living. The research described here takes a novel approach and uses data from a medical examiner office to create a database that can be used to improve the health of the living. Medical examiner data is a great resource for informaticians, even in its unstandardized form, because it includes several data elements absent in traditional medical records. The most important of these elements is cause of death as it is not always available in the electronic health record; however, medical examiner records also include autopsy reports, toxicology information, demographic variables, police reports, medical diagnoses, and medication lists.[5]

In New Mexico, The Office of the Medical Investigator (OMI) is a state-wide, centralized medical examiner's office. Any individual who dies in a sudden, violent, untimely, or unexpected manner, and any person who is found dead and the cause of death is unknown, is routed to the OMI for a possible autopsy (OMI website). The Center for Forensic Imaging at the OMI was awarded a grant from the National Institute of Justice in 2010 to evaluate whether postmortem computed tomography (CT) scans could supplement or supplant a traditional autopsy (2012-DN-BX-K019). Between 2010 and 2017, 85% of decedents that underwent an autopsy received a high resolution, full-body CT scan. This produced over 15,000 whole-body 3-D CT images. Each CT data set consists of two sets of images, optimized for soft tissue and bone. Each image set consists of 4,000 axial image slices, each with a 512 x 512 matrix and a slice thickness of 1 mm, with 0.5 mm overlap. In addition, each individual's record includes six scout images comparable to whole body radiographs of the decedent. However, records do not associate the scans with any organized lifestyle, health, or cause of death metadata.[7]

In 2014, a study determined the Minimum Data Set to associate with these CT images using a modified Delphi method. Researchers from a wide variety of fields (anthropology, medicine, forensics, informatics, epidemiology, biomedical research and dentistry), selected 59 variables through an iterative process that they believed to be essential to making the CT scans useful to researchers in multiple fields.[7] In 2016, the National Institute of Justice awarded a grant (2016-DN-BX-0144) to create the CT database with associated lifestyle, health, and cause of death information. The associated information includes nine additional variables that derive from the investigators' own research (grandparent's origins, marital status, ethnicity, and Hispanic identification). The data for all 68 variables derive from both VAST (the OMI's primary database that stores and organizes information regarding the cause of death, lifestyle and health information, in order to investigate the cause and manner of death), and through phone interviews with next of kin.[8-9]

Data in VAST are primarily in free, unstandardized text fields and requires natural language processing to extract. An example of the lack of standardization is the field sex/gender, which can be listed as M, m, male, or Male. The lack of standardization limits the ability to retrieve information effectively and requires much more time to clean before it is used. In its current form, as data complexity increases, data recovery becomes less accurate. In addition, information such as medications taken by the decedent are found in long free text fields. These free text fields require data extraction to make the data useful.[10]

We sought to extract, clean and standardize data from VAST to populate the 68 metadata variables in the new database in order to make the database useful to the widest array of researchers. The New Mexico Decedent Image Database (NMDID) will be available to bona fide researchers and educators at NMDID.unm.edu in early 2020.

Methods

Our first step was to alter the metadata to be HIPAA compliant in regards to personal health information (PHI), even though, according to NM state laws, the data are in the public domain (Inspection of Public Records Law). This step was undertaken as the investigators hope to expand the database to include more African American and Asian American decedents from other states. Other states may have stricter laws in place regarding the data from medical examiner's offices.[11] The 18 elements of PHI were considered with each metadata field, in particular this effected date of birth, date of death, and zip code. In addition, physical safeguards were installed in the database, with audit logs and limiting access through accounts.

Before any data standards could be applied, the 59 metadata fields, that were previously selected through a modified Delphi method with experts from multiple fields, had to be operationalized.[7] For example, the field "sex/gender" needed to be separated into two distinct fields, as it included two different aspects of data. Additionally, fields like "substance abuse" and "history of substance abuse" were grouped together so they could utilize the same normative answer list. The process of determining the content of each field was repeated for the entire metadata set.

Once the metadata fields were operationalized, each field was investigated using the Unified Medical Language System (UMLS) to determine relevant pre-existing data and vocabulary standards (see figure 1). In order to accomplish this, each term was searched in the UMLS using the metadata term, synonyms, medical and common language terms. The list of resulting standards was compiled and evaluated for effectiveness in this particular database by the authors. For example, some data were to be provided by next of kin, so could not be expected to be as technical as data coming from VAST. Vocabulary standards were then selected by the authors, modifications made if necessary, and implemented into the database. For those metadata fields where no known vocabulary standard or no normative answer list existed, the authors collaborated and determined a new standard for this particular data set (e.g. "alcohol use").

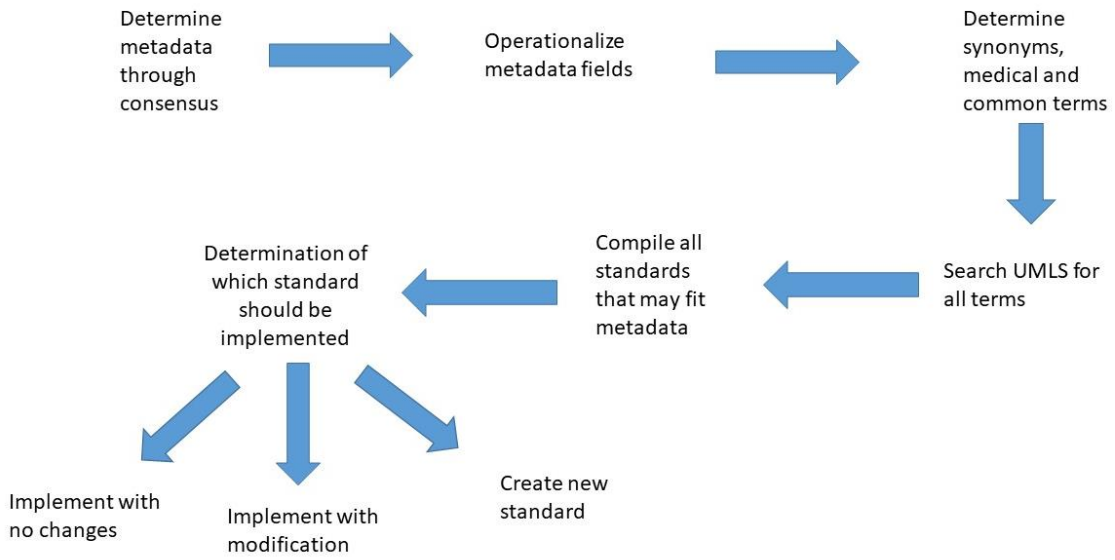

Figure 1. Method of metadata and vocabulary standard selection

The actual data was exported into Excel sheets from the relational database (VAST) in which it resides. Multiple fields required data cleaning prior to any analysis. This was accomplished using Open Refine.[12] In addition, may of the fields within VAST are completely free text fields, with no standardization except for how the data is entered. Data were extracted from free text fields and autopsy reports using Canary Natural Language Processing Software.[13] Canary is a data extraction freeware tool. Natural language processing was used to capture "medications", "cadaver length", "cadaver weight", and "environmental conditions" (i.e. where the decedent was found). In the future, we will add data for cases where next of kin could not be contacted by using Canary to extract "marital status," "socioeconomic status," "diagnoses," "activities," "drinking status," "drug use," and "occupations".

Results

NMDID, the new database to house the health and lifestyle data and scout images (preliminary image similar to an X-ray), was created in MySQL and currently contains 20 MB of data (without scout images). The CT scans are housed separately as to minimize the size of the database.

In order to remain consistent with HIPAA PHI standards, certain alterations were made to the metadata fields. For example, "birth date" and "death date" became year only, and "zip code" became only the first three digits. The data from the medical examiner are public and open to request in New Mexico, but may not be in other states. Because we might want to expand to include data from other states in the future, we designed the database with the most restrictive standards that still allowed data use.

After determining which variables needed to be combined, which variables should be separated into two or more fields, and adding additional variables specific to the authors' research, we had 68 metadata variables in three areas: census, health, and circumstances of death (see table 1).

For each of these metadata fields, there were three different ways a vocabulary standard could be selected/created:

1) There was an existing data standard that required no modification;
2) A data standard existed but required modifications;
3) No appropriate data standard or normative answer list existed, so a new standard was created.

Table 1. Final set of metadata that are HIPAA compliant and operationalized.

CENSUS	HEALTH	CIRCUMSTANCES OF DEATH
Sex	Cancer	Primary cause of death
Birth year	Congenital and genetic disorders	Contributing cause of death
Death year	Chromosomal abnormalities	Manner of death
Zip code	Dental health as an adult	Identification method
Age in years	Dental health as a child	Bone density
Age in months	Scoliosis	Death city
Gender	History of broken bones	Death county/tribal land
Number of pregnancies	Facial trauma	Cadaver condition
Number of live births	Plastic surgery	Cadaver length
Marital status	Surgery	Cadaver weight
Living weight	Implants	CT scan settings
Living height	Radiation therapy	Time delay after death
Race	Medical diagnoses	Name of person entering data
Tribal affiliation	Notes on medical diagnoses	
Ethnicity	Medications	
Hispanic identification	Substance usage	
Birth weight	Quit date of substance	
Birth weight category	Years of substance usage	
Birthplace	Tobacco type	
Years in the US	Tobacco usage	
Mother's birthplace	Drinking status	
Father's birthplace	Dietary pattern	
Mother's mother birthplace	Occupations	
Mother's father birthplace	Duration of occupations	
Father's mother birthplace	Activities	
Father's father birthplace	Strenuous lifting	
	Educational level	
	Socio-economic status as a child	
	Socio-economic status as an adult	
	Carcinogens	

Data extraction and cleaning

Once data standards had been determined, some data could be extracted from VAST to populate the database prior to calling next of kin. This information included standardized data (e.g. manner and cause of death), free text simple fields (e.g. method of identification, sex, race, ethnicity). The standardized data was imported directly into NMDID, while the simple free text fields were cleaned using OpenRefine and then imported.[12] A great deal of effort was spent extracting data from long, complex free text fields (e.g. scene investigation notes, medication notes; see table 2). Once the long free text fields were exported from VAST into Excel or Word documents, the needed data was extracted from the fields using Canary.[13] Using this software, code was developed that separately extracted two sets of data: case number and a medication list, and case number and "cadaver length" and "cadaver weight."[10] The "medication" lists were further altered to change them into RxNorm CUIs.[14] Additionally, the "cadaver weights" and "cadaver lengths" were in pounds and inches for half of the data and had to be converted to kilograms and centimeters.

The medications were extracted from a field that only contained medication data, and was standardized in how they were listed. Canary contains a list of medication names that can be used, however, it is not maintained and therefore does not necessarily contain all possible medications at the time of extraction. It was decided that the list was adequate as the medication list is verified when calling next of kin for additional data. “cadaver weight” and “cadaver heights” were extracted from the autopsy reports. Within the report many organ weights and lengths are given, thereby making it impossible to just extract measurements alone.

Discussion

Data standard without modification

Twenty-five of the 68 metadata fields had existing standards that could be applied without modification. Standards were derived from multiple sources, including LOINC (incorporating trial codes), SNOMED CT, internal standards specific to the OMI, ISCO, ICD-10 and RxNorm.[14-19]

One example of a metadata field that did not require any modifications was “cause of death”. The OMI had internal standards that were determined when the office originally switched from paper records to an electronic system. These codes represent five sections (unnatural, natural, other, pending and undetermined), which are each subdivided further. For example, unnatural includes carbon monoxide poising, stab wounds, and hanging, while natural includes pneumonia, obesity, and renal failure. There are a total of 83 subcategories that the pathologist may select for cause of death. As this is an internal standard that is unique to the OMI, in the future it will require a crosswalk between it and other systems.

Another, more common standard that did not require any modifications was “smoking status.” Since some of the data in investigation reports come from decedents’ medical records, we determined that this would be the best standard to follow. After examining all the standards available, we selected LOINC 72166-2 and SNOMED CT codes for the normative answer list, which are used in the electronic health record.[16-17]

Data standard with modifications

Seventeen of the 68 metadata fields required modification from an existing standard. These standards were primarily from LOINC (including trial codes).[15-16]

One example of a metadata field that required some modification was “birthplace.” LOINC code 63490-7 is a trial code from PHEN-X (consensus measures for Phenotypes and eXposures) that asks where patients were born.[16] The two possible answers are:

1) In the US, and
2) Outside the US.

Since our population includes many Mexican Americans and Mexican nationals, Mexico was added as an option for birthplace. We also added a category to account for cases of missing information. The result was the following normative answer list:

1) In the US,
2) In Mexico,
3) Outside US and Mexico, and
4) Unknown.

A second example is data describing decedent “drug use”. We contacted a substance abuse counselor to determine which standards were currently being used in practice. This included ICD-10 and DSM-5.[19-20] The ICD-10 categories were selected since DSM-5 has a category for alcohol, which is being recorded separately in our database. ICD-10 codes are for mental and behavioral disorders for specific drugs (F10-F19). F19 was modified to exclude an answer

of "multiple drugs," as the database is designed to record use of more than one drug type per individual. Additionally, each category was repeated to describe past drug use, as this is being captured in the same field (as it is for alcohol and smoking).

New data standard creation

The remaining 26 metadata fields could not be related to an appropriate normative, standard answer list. For example, "medical diagnoses" was a complicated field as information came from VAST and next of kin interviews. As a result, the data could be specific or vague. For this field, ICD-10 codes at the category level or precise SNOMED CT codes could have been implemented.[17,19] SNOMED CT standard is used to identify medical problems within electronic health records, while ICD-10 is used for billing.[17,19] Both standards require coding the response as a computer-readable number. Due to the vast number of precise SNOMED CT codes and the lack of everyday use of them by physicians and researchers, we excluded its use for this database. Category level ICD-10 codes were next considered, but there are a large number of ICD-10 codes that would likely not be found in our population (such as Pinta and Yaws). Further, next of kin may also not know the level of detail used in either ICD-10 or precise SNOMED CT codes regarding most diseases a decedent had experienced. As a result, a modified and more general approach has been undertaken. We generated a list of the 20 most common diseases, disorders, and conditions found in this population (through an informal survey with OMI pathologists) and their corresponding SNOMED CT codes. The result will be to combine some of the diagnoses data into one standard, but such a summary approach was deemed necessary given that the information can come from non-medical individuals (next of kin), yielding information that privileges accuracy over precision. A free text field has also been added to capture additional diagnoses that were not included. This allows for future addition of specific disease information to the database.

Drinking status was another field for which there was no appropriate data standard for the normative answer list. SNOMED CT contains a code for recording alcohol consumption and contains a normative answer list; however, it uses "recommended sensible amount" as the differentiator between normal and excessive use.[17] The National Institute on Alcohol Abuse and Alcoholism (NIAAA) considers "at-risk" drinking as more than 14 drinks per week for men and seven per week for women.[21] Therefore, the standard was created with these limits in mind:

1) Never drank,
2) Low risk (<14 drinks per week for men and <7 drinks per week for women),
3) High risk (>14 drinks per week for men and >7 drinks per week for women),
4) Previous high risk,
5) Drinker, current status unknown,
6) Unknown if ever drank.

The last two categories (5 and 6) were added in as next of kin may not know and the OMI's database may not contain the relevant information.

Data extraction and cleaning

Using Canary for data extraction came with some limitations.[13] When extracting medications from the free text field, many had been spelled incorrectly in the investigator's original notes. Others medication names were newer and not included in the medication list built into Canary. Information could not be extracted from the free text in either of these cases. This could have been rectified by updating the list. Since each medication is verified when calling next of kin, this additional step was not taken.

When extracting lengths and weights of the cadaver, some cases were missed because the data was not originally entered in the standard format. In general, the relevant text was formatted as "weighs XX kg, XX cm in length." This simplified extraction of the values and units from the external examination section of the autopsy report. It was not possible to search solely by number and units as multiple organ weights and lengths are reported in the autopsy report. A separate data extraction was run to capture infant or fetus lengths and weights as the values are standardized differently than adults. In addition, the case number was pulled from the report. However, if the report contained a number from another decedent, which might happen when two decedents were found together, Canary would on

occasion assign a cadaver length and weight to multiple decedents. However, this resulted in duplicates of the additional case number. To remedy this, data were searched for duplicates and corrected by hand. This was the case in 20 cases out of 15,248.

Table 2. Example of complex free text fields, de-identified, before data extraction (including drug misspellings).

Decedent 1: field supplemental notes	* Prescribed To: XX Presc #: X Pharmacy: MD/DO Prescribing: X, MD Medication Prescribed: Prednisone Dosage: 50mg Date Filled X Qty Issued: X Qty Left: X Qty and Frequency Taken: take 25 by once each day for 3 days only * Prescribed To: XX Prescription #: X Pharmacy: MD/DO Prescribing: X, MD Medication Prescribed: Betasron Dosage: 3mg Date Filled X Qty Issued: X Qty Left: X Qty and Frequency Taken: * Prescribed To: XX Prescription #: X Pharmacy: MD/DO Prescribing X, MD Medication Prescribed: Provigil Dosage: 200mg Date Filled X Qty Issued: X Qty Left: X Qty and Frequency Taken: 1 every day * Prescribed To: XX Prescription #: X Pharmacy: MD/DO Prescribing: X, MD Medication Prescribed: Detrol LA Dosage: 4mg Date Filled X Qty Issued: X Qty Left: X Qty and Frequency Taken: 1 monthly

Conclusion

Vocabulary standards and terminologies are of the utmost importance in the creation of databases, in order to ensure interoperability. For this reason, we searched medicine, forensics, anthropology, and other fields for standards to apply to our new database. Thirty-seven percent of the metadata fields could be mapped directly to existing standards. This included fields such as smoking status, sex, gender, and marital status. Twenty-five percent required a modification of the standard to apply to this particular database. Among those modified was birthplace and drug usage. The majority of the metadata fields (38%) required new standards to be developed or normative answer lists to be developed. One field was added as a free text field ("other medical diagnoses").

Medical examiner data is an underused resource for public health data, because, unlike vital statistics data, it has generally been unavailable to the research public. Medico-legal data has several advantages, including that it is associated with lifestyle and health data, unlike vital statistics, which, while easily available and standardized, is not very rich. The NMDID will provide 15,248 CT scans and associated metadata to researchers, creating an entirely new resource. NMDID sets a precedent, and perhaps some new standards, for the use of medico-legal data from other resources worldwide. NMDID will be available to researchers at the beginning of 2020 at NMDID.unm.edu.

Acknowledgements

Funded by National Institute of Justice 2016-DN-BX-0144. Statements made are solely the responsibility of the authors.

References

1. Parrish G. Assessing and improving the quality of data from medical examiners and coroners. In: Proceedings of the international collaborative effort on injury statistics 1995: 1(25):1-10.

2. Hanzlick R, Parrish RG. The role of medical examiners and coroners in public health surveillance and epidemiologic research. Annual Review of Public Health. 1996 May;17(1):383-409.
3. OMI, Office of the Medical Investigator 2010 Annual Report, University of New Mexico, 2010.
4. Graitcer PL, Williams WW, Finton RJ, Goodman RA, Thacker SB, Hanzlick R. An evaluation of the use of medical examiner data for epidemiologic surveillance. American journal of public health. 1987 Sep;77(9):1212-4.
5. Hanzlick R. Medical examiners, coroners, and public health: a review and update. Archives of pathology & laboratory medicine. 2006 Sep;130(9):1274-82.
6. Sinard JH, Powell SZ, Karcher DS. Pathology training in informatics: evolving to meet a growing need. Archives of Pathology and Laboratory Medicine. 2014 Apr;138(4):505-11.
7. S Berry, Metadata Determination for a Cadaveric Collection. Master [thesis], Albuquerque: University of New Mexico, 2014.
8. Berry SD, Edgar HJ. Development of a large-scale, whole body CT image databsase. In: AMIA Annual Symposium Proceedings 2017: 1951.
9. Edgar HJ, Berry SR. NMDID: A new research resource for biological anthropology. In: American Journal of physical Anthropology Supplemental 2019 Mar:168(S68): 66.
10. Berry SD, Edgar HJ. Research from records: retrieving and sharing useful data from a non-research database. In: American Journal of physical Anthropology Supplemental 2019 Mar:168(S68): 19.
11. Nelkin D, Andrews L. Do the dead have interests-Policy issues for research after life. Am. JL & Med.. 1998;24:261.
12. OpenRefine [software]. Available from http://openrefine.org/
13. Malmasi S, Sandor NL, Hosomura N, Goldberg M, Skentzos S, Turchin A. Canary: An NLP platform for clinicians and researchers. Applied clinical informatics. 2017 Apr;8(02):447-53.
14. RxNORM Available from https://www.nlm.nih.gov/research/umls/rxnorm/index.html
15. Forrey AW, Mcdonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, Fiers T, Charles L, Griffin B, Stalling F, Tullis A. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clinical chemistry. 1996 Jan 1;42(1):81-90.
16. LOINC. Available from https://loinc.org/
17. SNOMED CT. Available from https://www.nlm.nih.gov/healthit/snomedct/index.html
18. International Standard Classification of Occupations. Available from https://www.ilo.org/public/english/bureau/stat/isco/
19. ICD 10 CM. Available from https://www.cdc.gov/nchs/icd/icd10cm.htm
20. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub; 2013 May 22.
21. National Institute on Alcohol Abuse and Alcoholism, How much is too much?, Available from https://www.rethinkingdrinking.niaaa.nih.gov/How-much-is-too-much/Is-your-drinking-pattern-risky/Whats-At-Risk-Or-Heavy-Drinking.aspx.