Developing the Minimum Dataset for the New Mexico Decedent Image Database

Shamsi Daneshvari Berry[1,2] Philip J. Kroth[1] Heather J. H. Edgar[2] Teddy D. Warner[2]

[1] Western Michigan University, Homer Stryker M.D. School of Medicine, Kalamazoo, Michigan, United States
[2] University of New Mexico, Albuquerque, New Mexico, United States

Address for correspondence Shamsi Daneshvari Berry, PhD, MS, CPHI, 1000 Oakland Drive, Kalamazoo, MI 49008, United States (e-mail: shamsi.berry@med.wmich.edu).

Appl Clin Inform 2021;12:518–527.

Abstract

Background A minimum dataset (MDS) can be determined ad hoc by an investigator or small team; by a metadata expert; or by using a consensus method to take advantage of the global knowledge and expertise of a large group of experts. The first method is the most commonly applied.

Objective Here, we describe a use of the third approach using a modified Delphi method to determine the optimal MDS for a dataset of full body computed tomography scans. The scans are of decedents whose deaths were investigated at the New Mexico Office of the Medical Investigator and constitute the New Mexico Decedent Image Database (NMDID).

Methods The authors initiated the consensus process by suggesting 50 original variables to elicit expert reactions. Experts were recruited from a variety of scientific disciplines and from around the world. Three rounds of variable selection showed high rates of consensus.

Results In total, 59 variables were selected, only 52% of which the original resource authors selected. Using a snowball method, a second set of experts was recruited to validate the variables chosen in the design phase. During the validation phase, no variables were selected for deletion.

Conclusion NMDID is likely to remain more "future proof" than if a single metadata expert or only the original team of investigators designed the metadata.

Keywords
- minimum dataset
- metadata
- future uses
- Delphi method
- snowball sampling

Introduction

The total amount of digitally stored data now exceeds 44 zetabytes.[1] These data are comprised of both data collected for scientific inquiry and those created as an artifact of another nonresearch purpose. For most research databases, the goal in their creation is to gather information regarding a scientific query, investigation, or task immediately at hand.[2,3] Data collected for other purposes, such as clinical work on the other hand, are used as artifacts of their original purpose. Regardless of whether data are collected for a specific purpose or exist as artifacts, the database is highly dependent on how well the dataset designers predicted its future uses, making it more "future proof," by selection of optimal metadata. Unfortunately, many selected database variables tend to be chosen to support immediate project needs and usually are chosen without an eye to future applications.

Image databases are growing in number and size, with differing associated modalities and variables. Multiple image databases are available in the health field, such as MedPix and the Cancer Imaging Archive. The cases available are organ specific, and the data are related to disease.[4,5] Because

received
January 25, 2021
accepted after revision
April 30, 2021

DOI https://doi.org/10.1055/s-0041-1730999.
ISSN 1869-0327.

the technology to search images per se is not yet widely available or standardized, a second associated database is needed to contain metadata associated with each image.[6] Users search on the images' meta-database (or, in this case, the data about the image data) to find images of interest.[7] The overall selection of metadata variables influences the breadth and variety of research that can be conducted. The quality of the set of metadata are inextricably linked to the quality and ease of the process of gathering, selecting, and transforming the data to answer an analytical question and therefore determine the value of the data in the future. However, it is difficult to predict all potential future uses of a database and so also difficult to predict the best metadata fields to select at the outset.

How can database designers determine metadata in a way that optimizes the value of the database for future users? In a world with unlimited time and funding, unlimited possible metadata fields would be desired. However, resource constraints limit the sophistication of the metadata design to a relatively small set of variables deemed most important at the time of initial design. Selecting too few or inappropriate variables can significantly reduce the value of the data over time by limiting the potential for reuse of the data. In contrast, defining too many fields requires increased use of valuable resources and reaches a point of diminishing returns. The challenge in database design is to define a reasonably sized meta-dataset that will produce high value now and in the future. The process of designing highly useful minimal datasets is critical to maximize the value of research data over time. As a result, how and by whom the metadata are selected affects the usefulness of the database, both now and in the future.

Metadata

Metadata are the structured information that characterizes each case in the primary database and supports additional functions or actions about an object, topic, or person.[7] Good quality metadata allows the user to efficiently retrieve information in a timely manner, whereas poor quality metadata may miss pertinent cases in a database.[8] The use of appropriate and high quality metadata facilitates information retrieval, searching, maintenance, understanding, interoperability, and reuse.[9,10]

In this current technology heavy world with expanding data, there are frequent opportunities for data collection and "data wrangling." For example, images for medical purposes are occurring every day in hospitals, doctor's offices, imaging facilities, and coroner/medical examiner offices. In 2019, an estimated 91 million computed tomography (CT) scans were performed in the United States alone.[11] At present, the majority of these images are stored in picture archiving and communication system,[12] encoded with Digital Imaging and Communications in Medicine standards,[13] and the Abbreviated Injury Scale,[14] but without easily linked health-related metadata.[15,16] Without association of such primary information with the other metadata, the ability to facilitate research or reuse data are limited. As the number of images created and stored continues to grow, few facilities are incorporating plans for image reuse by investigators and educators.

Metadata Selection

The effectiveness of retrieving data is dependent upon the number of metadata fields and the content contained within. It is a balance of discovery and cost, where additional information is available with more metadata fields, but costs more resources.[10] New metadata fields can be added as necessary, making the database more adaptable (add, delete, and change variables). However, it requires significant resources to back fill the data variables added.[10]

Minimum Dataset Creation

Individual metadata elements can be combined to form a set of data for an image or object, called a minimum dataset (MDS).[17] A MDS allows for interoperability of data between investigators in the healthcare system and research domains.[18–23] Major domains using an MDS to standardize retrieval of vital information include nursing,[23,24] genetics,[25] nursing homes,[26] spine trauma,[20] Infertility registry,[21] autoimmune disorders,[22] brain injury,[19] and studies of rare and orphaned diseases.[27,28]

Multiple approaches can be undertaken to select metadata, including through the resource author (those conducting the research or collecting the data—usually the most common approach), a metadata specialist, or a collaborative procedure.[8,19–25,27,28] Evidence has shown that many resource authors lack the skills and training to index or apply terminology standards and theories. Therefore, they often create metadata that is insufficient for conducting their research or any research beyond their immediate needs. Inadequate metadata weakens the ability to discover relevant records and can produce underpowered results. Using a metadata specialist can have the same problems, as they lack knowledge about the specific science being undertaken.[29] Greenberg and Robertson[29] suggest that the best quality metadata are obtained through a collaborative process. The exact method for collaboration can vary depending on the resources available for the creation of the MDS. The methods can include the Delphi method, in which there is no direct interaction, and the Nominal Group Technique, in which a round-robin discussion occurs.[30]

Assessment of a Minimum Dataset

To ensure its potential use beyond the immediate research purpose, the quality of the MDS should be evaluated. A variety of assessments have been used in the past to evaluate MDS; therefore, the evaluation of a MDS is not a consistent practice. However, the most common procedure is a survey.[31]

Objectives

The Office of the Medical Investigator (OMI) is the centralized medical examiner's office for the State of New Mexico. Medical examiner cases are thought to primarily be from homicide or suicide deaths. However, the vast majority of

Table 1 Manner of death at the old myocardial infarction in 2010 and 2017, and the new Mexico decedent image database (mid-2010 to mid-2017)[32,34]

Manner of Death	2010	2017	NMDID
Natural	24.7%	27.4%	34.6%
Accidental	35.4%	40.8%	38.6%
Suicides	16.8%	12.9%	15.4%
Homicides	9.5%	12%	7.4%
Undetermined or pending	13.5%	5.1%	4%[a]

Abbreviation: NMDID, New Mexico Decedent Image Database.
[a]Undetermined only.

cases in 2010 and in 2017 were from natural or accidental causes (►Table 1).[32] In addition, the autopsied OMI sample consist of the ethnic and racial composition of the state. In the 2010 census, 49% reported as Hispanic and 11% as Native American.[33] For the OMI sample, 30% were Hispanic in 2010 and 29% in 2017. Native Americans accounted for 9% of deaths routed to the OMI in 2010 and 2017.[32,34]

The Center for Forensic Imaging at the OMI was awarded in 2010 a grant from the National Institute of Justice to evaluate the efficacy of postmortem computed tomography (CT) scans to supplement or supplant a traditional autopsy (2010-DN-BX-K205). As a result, roughly 85% of decedents who underwent an autopsy at the OMI received a high resolution, head-to-toe CT scan. This produced thousands of whole-body 3D CT images between 2010 and 2017—a treasure trove for a variety of research domains—but with no organized and associated metadata to allow investigators to efficiently identify images of interest. As with the vast amount of data in healthcare, curation of the OMI dataset for both education and research is greatly needed.[18]

The OMI collected data for nonresearch purposes, that is, investigation, similar to the biomedical field, and healthcare data. These data would be lacking completeness and breadth needed for the effective use of the images. For this reason, we determined to collect additional data in interviews with next of kin.

The incorporation of a comprehensive annotation schema into a database occurred with the creation of the New Mexico Decedent Image Database (NMDID). NMDID facilitates future research using the CT images and associated health and lifestyle information by making them efficiently findable. NMDID is a unique resource due to its size, 3D images, and diverse population.

To design an optimal MDS, we used a collaborative procedure to choose the metadata for NMDID. We had two objectives:

- To determine the MDS to associate with CT scans in a database of 3D, whole-body, decedent images developed at the OMI. The MDS should enable investigators, from multiple domains, to efficiently and effectively search for images from the database that meet the inclusion and exclusion criteria of their studies with optimal sensitivity and specificity.
- To assess the relevance of the selected metadata. The MDS should be validated by a second separate group of experts to verify its usefulness in conducting research.

Methods

Design

We selected a consensus method to create the MDS to reduce biases from wither a single database creator or metadata specialist.[29] Furthermore, we used an electronic version to avoid the costs of an in person meeting. Electronic consensus was also asynchronous, so that each participant could do the work at their convenience. We chose the Delphi method because it facilitated electronic data collection[35]; however, other consensus methods would have also been appropriate. The Delphi method involves asking experts from relevant domains to obtain convergence of opinion.[36] The method allows for anonymous participation of experts through an iterative process. Due to the varying nature of each consensus panel, the level of consensus should be determined after each round to determine when additional rounds are no longer needed.

Once an MDS is determined through an iterative process, it needs to be validated or assessed by additional experts not involved with its creation, to ensure objectivity.[31,37] Questionnaires are regularly used to validate an MDS. The process outlined here did not specify how the selected fields will be collected or coded with appropriate standards. Encoding of the metadata took place after the MDS was defined.[38] The methods used in design and validation phases are illustrated in ►Fig. 1.

Expert Determination

For the design phase, we formulated a list of domains amenable to using whole-body, 3D, cadaveric CT scans, and the associated data for future research. Within each of the domains, peer-reviewed literature was searched to find experts. In addition, each of the authors suggested experts within their respective disciplines as well as individuals to contact for their recommendations.

For the validation phase, the design phase participants were asked to recommend two to three experts[39] whom they believed might use the full-body, 3D, cadaveric CT scans, and associated database with health, lifestyle, demographic and cause of death data. The design phase participants were asked for name, institution, and email address (if known) for each validation expert they recommended. This question was included in the third round of the Delphi survey.

All participants in the surveys had terminal degrees in their fields (MD, PhD, RN). Additional data on the participants were not captured; however, most were based in the United States with a smaller percentage being international.

Questionnaire Creation

A preliminary questionnaire was built in REDCap,[40] a web-based, open-source data capture program that has security

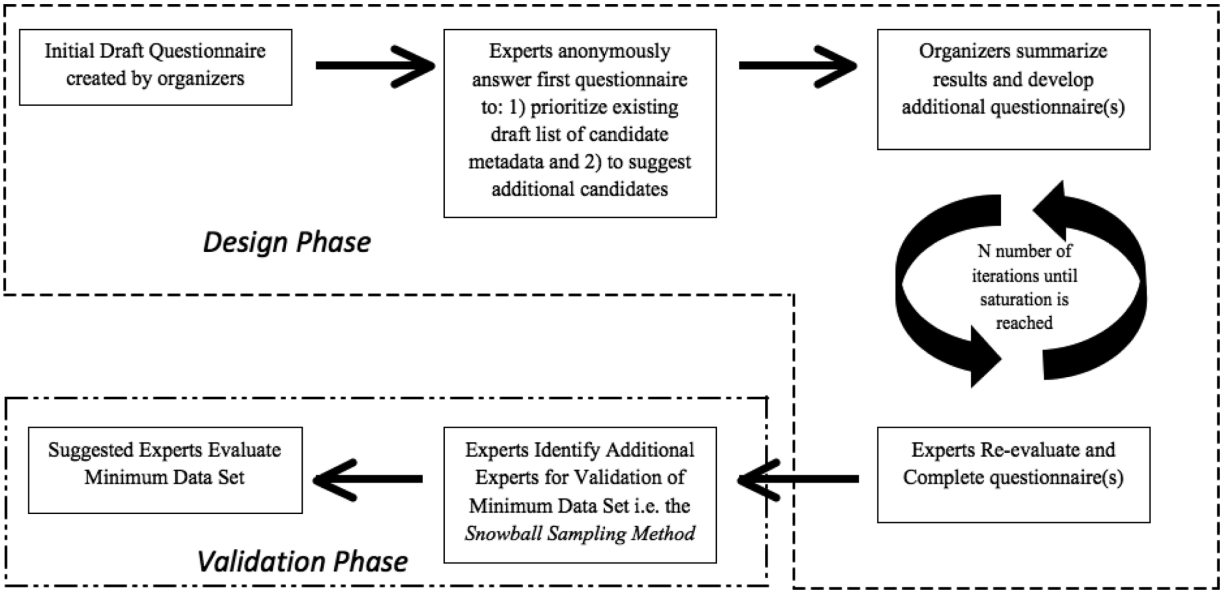

Fig. 1 Methods for designing (Delphi) and validating (snowball expert sampling) a minimum dataset.

and privacy controls. The initial questionnaire consisted of variables that had been suggested by the resource authors (S.D.B., H.J.H.E., P.J.K.) as important fields to include in the MDS. As such, these initial variables could be proxies for a resource author's database fields. The first questionnaire provided a basic set of variables within five categories: (1) personal characteristics, (2) lifestyle, (3) health, (4) occupation, and (5) other variables. For all rounds of the design phase, experts voted on original fields and the additional fields they suggested. When original terms were selected as not important for the database, they were eliminated.

The follow-up questionnaires in the design phase allowed participants to revise the groups and their own ideas. This process continued until we believed saturation had been reached. The last questionnaire within the design phase also asked participants to rate the suggested database fields in terms of importance of inclusion in the MDS (e.g., from 0 = not important at all to 10 = absolutely essential to include).

The REDCap validation questionnaire asked participants to evaluate the database fields and rate them. In addition, the experts were asked to provide any essential fields that the design phase participants did not identify.

For both the design and validation phases, a one-page recruitment letter was mailed to potential experts, as well as sent electronically to their institution email; that letter also included a one-page consent form. Because this project collected only nonsensitive data, we requested and received a waiver for a signed documentation of informed consent.

Results

Design Phase

A total of 72 experts were sent a letter and email asking for participation in the design phase. The 17 domains surveyed are listed in ►Table 2. In total, 42 participants (58% response rate) completed the questionnaire. Thirty-two experts self-identified their research domain; the summary is listed in ►Table 3. The emails and letters were sent at the end of September to coincide with the beginning of the fall school schedule. The questionnaire remained open until the end of November (10 weeks total).

The first questionnaire contained 50 original database variables (see ►Appendix A for list) for experts to evaluate and discuss. If a variable was suggested for elimination by the expert, they were asked to provide a reason. At the end of each section experts were asked what additional database variables they advised to have included. This included an "other" category where variables that were outside the five categories could be suggested. Consensus was defined as 60% agreement. In round 1, only four variables were eliminated from the list: last name, first name, marital status, and current residence address.

One investigator (S.D.B.) summarized the results and combined similar suggestions. The second questionnaire contained 120 database variables (including 46 of the original variables) for experts to evaluate. Thirty-three participants (46% response rate) responded to round 2 of the design phase. Round 2 was completed in 2 weeks. Agreement on inclusion of the database variables was extremely high despite the variation in the experts' research domains. As a result, consensus for round 2 was defined as 93%. This value was selected due to a large number of tied variables for importance below 93% (with 50 variables at 92 to 80% consensus). This cut-off point resulted in a manageable number of variables since the data would be coming from calling next of kin and extraction from the medical examiner's database.[41] After elimination of database variables that had less than 93% consensus, one additional variable was added back in (normal height) since related variable (cadaveric height) was part of the MDS. A total of 59 database variables remained after round 2 of the Delphi method determination of the MDS (►Table 4 for the MDS).

Table 2 Research domains sent surveys

Research domains	
Informatics	16
Epidemiology	7
Anthropology	4
Forensic anthropology	4
Forensics	4
Dentistry	3
Growth and development	3
Medicine	3
Biomechanics	2
Demography	2
Health disparities	2
Health information exchanges	2
Imaging research	2
Odontology	2
Orthopedics	2
Pathology	2
Population variation	2
Public health	2
Radiology	2
Secular change	2
Chronic pain	1
Dental anthropology	1
Health economist	1
Missing person databases	1

Table 3 Research domains of participants in the design phase

Experts' self-identified research domains	Count
Forensic anthropology	7
Anthropology	4
Biomedical informatics	3
Clinical informatics	2
Forensic radiology	2
Biological anthropology	1
Cognitive neuroscience	1
Data management	1
Demography, anthropology	1
Dental medicine, forensic dentistry, paleodontology	1
Emergency medicine	1
Forensic odontology	1
Forensic pathology	1
Health economics and health services research	1
Health services research	1
Health services/epidemiology	1
Pediatric orthopedic surgery	1
Skeletal pathology	1
Unanswered	10
Total	42

Round three (31% participation) helped to determine the importance of the variables on a sliding scale from 0 (not important at all) to 10 (essential). This round allowed reduction in the number of variables to be collected if funding was not adequate to capture all 59 variables.

Validation

A true validation would require years of data usage on the database. As this was not possible when determining the metadata, an assessment of the fields was used. This assessment helped to determine how useful the MDS is to researchers outside of the design group.

A total of 34 experts were suggested by 15 design phase participants. Fifty-three percent of participants responded (►Table 5 for their self-identified primary field of interest), suggesting variables for elimination and rating the database fields in order of importance in the MDS.

No variables were selected for deletion from the MDS during the validation phase. The level of consensus was lower during this portion; however, the majority of variables had greater than 70% consensus (31/59 variables). This demonstrates that the variables selected by the design phase participants were thorough in the selection process and included roughly 60% of the variables the validation phase participants would need for their research.

The validation phase also allowed for additional variables not included in the design phase to be elucidated. Fourteen variables were suggested for addition by the validation phase participants, with only three variables not included or inferred from the original MDS: maxillofacial skeletal relationship category, dental occlusion category, and organ weights. Since most of the variables the researchers wanted were actually included in the database or could be inferred, the 60% estimate of variable usefulness is an understatement. ►Table 6 for the complete list of variables suggested.

Discussion

The future value of a research database is based on the quality of metadata and an optimal design of the MDS. This is especially true in the realm of image databases. Because the technology to search on the images themselves is in its infancy and not yet ubiquitous,[42] the discovery of specific images of interest relies heavily on the quality of the metadata design. Without sufficient metadata, images will be significantly less discoverable and the sensitivity and specificity of a search or query will decrease markedly.

Table 4 Fifty-nine metadata variables selected through a modified Delphi method

Health characteristics		Lifestyle characteristics	Personal characteristics	Highest educational level	Other characteristics
Birth weight	Family history of cancer	Repetitive or habitual activities	Date of birth	Childhood socioeconomic status	Primary cause of death
Congenital abnormalities	History of radiation therapy	Current smoking status	Date of death	Adult socioeconomic status	Contributing cause of death
Medical diagnoses	Facial trauma	Smoking history	zip code	**Occupational characteristics**	Manner of death
Surgical history	Implanted devises	Current drinking status	Sex/gender	Current occupation	Time delay between death and CT scan
Current medications	Genetic disorders	Drinking history	Race	Length at current occupation	Location of death
Current height	Scoliosis	Current drug use	Country of origin	Major occupation during life	Environmental conditions of cadaver
Cadaver length		Drug use history	Number of years in US	Occupation history	Method for identification
Current weight	Dental health as a child	Dietary pattern	Parents' country of origin	Exposure to carcinogens	CT scanner settings
Cadaver weight	Dental health as an adult		Number of pregnancies	Strenuous lifting	Person entering data
Bone density	Presence of dental caries		Number of live births	Length of military service	

Abbreviations: CT, computed tomography; US, United States.

Table 5 Self-identified research domains of participants in the validation phase

Experts' self-identified research domains	Count
Forensic anthropology	4
Anthropology	3
Biological anthropology	1
Dentistry	1
Forensic odontology	1
Forensic pathology	1
Interprofessional collaboration	1
Medical devices	1
Medical imaging	1
Modern human skeletal variation	1
Physical anthropology	1
Skeletal biology	1
Total	17

Therefore, appropriate metadata are vital, yet conceptually complicated, requiring a thoughtful balance between discoverability of relevant images and the resources necessary to design and construct a sufficient MDS. Using a consensus method with experts from varying domains is a valuable approach to improve the quality and completeness of the chosen variables and lessen bias.[29] Although varying domains were sought, in some fields experts were not be determined such as public safety. Additionally, the majority of respondents in both the design and validation phases could be classified as anthropologists. This could add bias to how "future proof" the database will become. However, the diversity of research within anthropology is great and those surveyed performed very different research. In addition, many domains not surveyed have used the database to date, such as public safety, art, and virtual education.

Although this method is robust in its ability to identify potentially "future proof" metadata, it is not infallible. Not all variables are discoverable, even after three rounds with experts suggesting and editing metadata fields, and a validation round in which additional participants recommended further variables. Researchers eliminated marital status as a variable in the first round, and it was not suggested for inclusion by the validation phase participants. This is surprising given that it is commonly included in health datasets as good indicator of health.[43–45]

After a consensus of 93% was imposed on round 2, 59 variables remained. For round 3, the experts were asked if a variable should be kept in the database and how important it might be to future research using the database. This provided us with the ability to create a sliding cut-off point depending on how many final fields we could include in the database.

Some of the variables chosen as important and those eliminated were surprising. The final list of database fields ($n = 59$) contained only 26 original variables (52%) of the 50

Table 6 Variables suggested in validation phase

Variables suggested for inclusion	MDS variable can be inferred from	MDS variable can be an additional response	Number of participants suggesting change
Absence/presence of removable dental implants		Implanted devices	1
Occupation of parents	Childhood socioeconomic status		1
Income of parents	Childhood socioeconomic status		1
Income of decedent	Adult socioeconomic status		1
Exercise habits	Habitual activity		1
How consistent was exercise	Habitual activity		1
Was the individual an athlete	Habitual activity		1
Presence of amputations		Major surgeries	1
Presence of surgical implants		Implanted devices	1
Trauma present at death	History of broken bones, primary cause of death, and contributing cause of death		2
Age	Date of death and date of birth		1
Maxillofacial skeletal category			1
Dental occlusion category			1
Organ weights			1

Abbreviation: MDS, minimum dataset.

selected by authors (S.D.B., H.J.H.E., P.J.K.). ►Table 7 summarizes the variable counts. The vast majority of final variables were suggested by the experts and validated by a separate group. The resulting fields spanned personal characteristics, circumstances of death, health, and lifestyle as well as CT settings. This process supports the value of a consensus method incorporating opinions beyond those of the current project designers.

The validation phase also demonstrated that more than 60% of the variables were of interest to the participants for their specific research questions. In addition, of those variables suggested for inclusion in the MDS by the validation group, only three could not be deemed equivalent to existing variables. This suggests that the method of development for this MDS was successful in its attempt to be more future proof and accommodate research from multiple domains.

NMDID became available to the research public in February 2020; as of November 3, 2020, there were 327 users representing 34 countries. The data and images have been used for research on multiple projects including biomechanics, COVID-19, traumatic injury analysis, dental development, art, sarcoidosis, Hispanic diversity, obesity research, and virtual education.[41] While we expected that education might constitute a relatively minor component of uses for NMDID, the requirements imposed by the COVID-19 pandemic significantly increased these applications, providing a case-study for the unexpected value of future proofing.

Table 7 Number of database variables by round in design phase

Round	Respondents (response rate)	Number of variables evaluated	Number of original variables	Consensus cut-off point
1	42 (68%)	50	50	60%
2	33 (46%)	120	47	93%
3	22 (31%)	59	26	NA

Abbreviation: NA, not applicable.

Conclusion

Using virtual Delphi and snowball methodologies to obtain consensus can be an extremely beneficial tool for MDS design. These two methods require a large number of experts to consider appropriate variables but can be conducted at a relatively low cost. Furthermore, by requiring consensus among disparate researchers, bias that may be inherent in one individual's metadata creation can be balanced by the opinions of others. Other consensus methods may also be beneficial but would also require the diverse domains queried and a validation phase.

It is difficult to ensure that any database will be "future proof." However, the database will likely remain more relevant in the future if more than a single metadata expert or original team of investigators designed the metadata. In this case, if only database creators (S.D.B., H.J.H.E., P.J.K.) had been consulted for MDS creation for NMDID, over 56% of final variables would not have been captured. This research suggests not only is expert group opinion the path to follow for MDS development, but diverse representation is vital for making a MDS more "future proof."

The variables such as operationalized, vocabulary standards applied, and seven additional fields of interest to the authors (including marital status) were added before NMDID was built. The operationalization phase required joining some variables together such as current and former smoking status, and breaking others apart such as sex and gender. In the final MDS, there are 69 variables.[46] The database is currently freely available at NMDID.UNM.EDU.

Clinical Relevance Statement

This article demonstrates a method to ensure a database is more "future proof" when created from the artifact of care.

Multiple Choice Questions

1. The best method for creating the metadata of a database is to query:
   a. The resource author
   b. A metadata specialist
   c. The consensus of experts

   Correct Answer: The answer is option c. A resource author is biased to their own research and a metadata specialist does not know the research topic as well. So, the best technique for lessening biases is to use a consensus of experts.[29]

2. A "future-proof" database allows for:
   a. Only the original research question to be answered
   b. Research beyond the original purpose
   c. Previous research to be reanalyzed

   Correct Answer: The correct answer is option a. A future-proof database allows for the original research question as well as research beyond the original purpose. It ensures that future questions can be answered.

Protection of Human and Animal Subjects

We received Institutional Review Board approval from the University of New Mexico Human Subjects Research Review Committee on June 10, 2013 (Human Research Protections Office13–229).

Funding

This study is funded by National Institute of Justice 2016-DN-BX-0144.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to thank the Office of the Medical Investigator in Albuquerque, NM. Statements made are solely the responsibility of the authors

References

1 How much data is created every day? [27 powerful stats]. Accessed November 24, 2020 at: https://seedscientific.com/how-much-data-is-created-every-day/
2 Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5(01): 180178
3 Bielefeld RA, Yamashita TS, Kerekes EF, Ercanli E, Singer LT. A research database for improved data management and analysis in longitudinal studies. MD Comput 1995;12(03):200–205
4 National Library of Medicine. MedPix. Accessed 2021 at: https://medpix.nlm.nih.gov/home
5 Cancer Imaging Archive. Accessed 2021 at: https://www.cancerimagingarchive.net/
6 Tagare HD, Jaffe CC, Duncan J. Medical image databases: a content-based retrieval approach. J Am Med Inform Assoc 1997;4(03):184–198
7 Greenberg J. Metadata generation: processes, people and tools. Bull Am Soc Inf Sci Technol 2005;29(02):16–19
8 Sarah C, Jane B, Rónán O, Ben R. Quality assurance for digital learning object repositories: issues for the metadata creation process. ALT J 2004;12(01):5–20
9 Sicilia M-Á Metadata, semantics, and ontology:providing meaning to information resources. Int J Metadata Semant Ontol 2006;1 (01):83–86
10 Malaxa V, Douglas I. A Framework for metadata creation tools. Interdiscip J E Learning Learn Objects 2005;1(01):151–162
11 2019 CT Market Outlook Report. Accessed 2019 at: https://imvinfo.com/product/2020-ct-market-outlook-report/
12 Choplin RH, Boehme JM II, Maynard CD. Picture archiving and communication systems: an overview. Radiographics 1992;12 (01):127–129
13 DICOM. Accessed Accessed November 24, 2020 at: https://www.dicomstandard.org/
14 Greenspan L, McLellan BA, Greig H. Abbreviated injury scale and injury severity score: a scoring chart. J Trauma 1985;25(01):60–64
15 Annamalai M, Guo D, Susan M, Sep JS. 2009 U. Oracle database 11 g DICOM medical image support. Accessed 2009 at: https://download.oracle.com/otndocs/products/multimedia/pdf/oow2009/mm_oow09_dicom_S311474.pdf

16 Health at a Glance 2017: OECD Indicators. Accessed 2017 at: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2017_health_glance-2017-en
17 Health Information Policy Council. Background paper: uniform minimum health data sets. Accessed 1983 at: https://link.springer.com/chapter/10.1007/978-1-4757-4160-5_18
18 Werley HH, Devine EC, Zorn CR, Ryan P, Westra BL. The nursing minimum data set: abstraction tool for standardized, comparable, essential data. Am J Public Health 1991;81(04):421–426
19 Domensino AF, Winkens I, van Haastregt JCM, van Bennekom CAM, van Heugten CM. Defining the content of a minimal dataset for acquired brain injury using a Delphi procedure. Health Qual Life Outcomes 2020;18(01):30
20 Tee JW, Chan CHP, Gruen RL, et al. Inception of an Australian Spine Trauma Registry: The Minimum Dataset. Accessed 2012 at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3864422/
21 Abbasi M, Ahmadian L, Amirian M, Tabesh H, Eslami S. The Development of a Minimum Data Set for an Infertility Registry. Perspect Heal Inf Manag; 2018
22 McCann LJ, Kirkham JJ, Wedderburn LR, et al. Development of an internationally agreed minimal dataset for juvenile dermatomyositis (JDM) for clinical and research use. Trials 2015;16(01):268
23 Ranegger R, Hackl WO, Ammenwerth E. A proposal for an Austrian nursing minimum data set (NMDS): a delphi study. Appl Clin Inform 2014;5(02):538–547
24 Werley HH, Lang NM, Westlake SK. Brief summary of the nursing minimum data set conference. Nurs Manage 1986;17(07):42–45
25 Meaney FJ, Cunningham GC, Riggle SM. Development of a national genetic services database. Proc Symp Comput Appl Med Care. Accessed 1991 at: Published online 1991:424–428https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2247567/
26 Porock D, Oliver DP, Zweig S, et al. Predicting death in the nursing home: development and validation of the 6-month minimum data set mortality risk index. J Gerontol A Biol Sci Med Sci 2005;60(04):491–498
27 Rubinstein YR, Groft SC, Bartek R, et al. Creating a global rare disease patient registry linked to a rare diseases biorepository database: Rare Disease-HUB (RD-HUB). Contemp Clin Trials 2010;31(05):394–404
28 Jenders R, McDonald C, Rubinstein Y, Groft S. Applying standards to public health: an information model for a global rare-diseases registry. Accessed 2011 at: Published online 2011: 1819https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900177/
29 Greenberg J, Robertson WD. Semantic Web Construction: An Inquiry of Authors' Views on Collaborative Metadata Generation. Vol 0.; 2002
30 Bagley Thompson C, Schaffer J. Minimum data set development: air transport time-related terms. Int J Med Inform 2002;65(02):121–133
31 Hillmann DI. Metadata quality: From evaluation to augmentation. Cat Classif Q 2008;46(01):65–80
32 Zumwalt RE, Aurelius M, Brooks E, et al. 2010 Annual report office of the medical investigator state of New Mexico. Accessed 2010 at: https://hsc.unm.edu/omi/_docs/pdfs/ar2010.pdf
33 2010 Census: new Mexico profile. Accessed August 5, 2020 at: https://www2.census.gov/geo/maps/dc10_thematic/2010_Profile/2010_Profile_Map_New_Mexico.pdf
34 New Mexico office of the medical investigator annual report. Accessed 2017 at: https://hsc.unm.edu/omi/_docs/pdfs/ar2018.pdf
35 Yousuf MI. Using experts'experts'opinions through Delphi technique. Pract Assess, Res Eval 2007;12(04):
36 Hsu C-C, Sandford BA. The Delphi technique: making sense of consensus. Pract Assess, Res Eval 2007;12:10
37 Goossen WTF, Epping PJMM, Feuth T, Dassen TWN, Hasman A, van den Heuvel WJA. A comparison of nursing minimal data sets. J Am Med Inform Assoc 1998;5(02):152–163
38 Berry SD, Edgar HJH. Standardizing data from the dead. Stud Health Technol Inform 2019;264:1427–1428
39 Leo A. Goodman. Snowball Sampling. Ann Math Stat 1961;32(01):148–170
40 Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(02):377–381
41 Edgar H, Daneshvari Berry S, Moes E, Adolphi N, Bridges P, Nolte K. New Mexico decedent image database. Office of the Medical InvestigatorUniversity of New Mexico2020
42 Hou J, Chen Z, Qin X, Zhang D. Automatic image search based on improved feature descriptors and decision tree. Integr Comput Aided Eng 2011;18(02):167–180
43 Robards J, Evandrou M, Falkingham J, Vlachantoni A. Marital status, health and mortality. Maturitas 2012;73(04):295–299
44 Verbrugge LM. Marital Status and Health. Vol 41.; Accessed 2021 at: https://psycnet.apa.org/record/1980-27843-001
45 Umberson D. Gender, marital status and the social control of health behavior. Soc Sci Med 1992;34(08):907–917
46 Berry SD, Edgar HJH. Extracting and standardizing medical examiner data to improve health. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits. Transl Sci 2020;2020:63–70

Appendix A Original variables

Personal characteristics	Lifestyle characteristics	Health characteristics	Occupational characteristics	Other characteristics
Last name	Hobbies	Medical diagnoses	Current occupation	Primary cause of death
First name	Current exercise status	Surgical history	Length at occupation	Secondary cause of death
Date of birth	Exercise history	Height	Occupation history	Medical insurance status
Date of death	Current smoking status	Current weight		
Current residence address	Smoking history	Weight history		
Length at current residence	Current drinking status	Childhood health status		
Marital status	Drinking history	Diabetes history		
Sex/gender	Current drug use status	Family history of cancer		
Race	Drug use history	Cancer diagnosis		
Hispanic ethnicity		High blood pressure history		
Country of origin		History of broken bones		
Parents' country of origin		History of other diseases/disorders		
Number of pregnancies		Dental health as a child		
Number of live births		Dental health as an adult		
Number of living offspring				
Annual income				
Highest education level				
Handedness				

