MAINTAINING PATIENT CONFIDENTIALITY IN THE PUBLIC DOMAIN INTERNET AUTOPSY DATABASE (IAD)

Jules J.Berman, Ph.D., M.D., G.William Moore, M.D., Ph.D., and Grover M.Hutchins, M.D.

Veterans Affairs Maryland Health Care System, University of Maryland School of Medicine, and The Johns Hopkins Medical Institutions, Baltimore, MD

The Internet provides the opportunity of permitting public access to large databases containing patient information that can be shared and utilized by epidemiologists, health planners, and medical researchers. Until now, large databases containing patient information have been held in strict confidence, with database access available only to approved researchers or to researchers with access limited to only specific portions of the database. The Internet Autopsy Database (IAD) consists of demographic and pathologic data from over 49,000 autopsies contributed by over a dozen academic medical institutions. Each autopsy record in the public database consists of a uniform set of demographics and SNOMED-compatible terms. To make the database publicly available, a strategy had to be devised that assured the privacy of every person included in the database. A key step involved translating the autopsy facesheets into a listing of SNOMED-compatible terms that effectively eliminated identifying terminology, replacing free text with a generic nomenclature that preserves diagnostic information. The entire database is available on the Internet at: http://www.med.jhu.edu/pathology/iad.html

INTRODUCTION

In 1975, the College of American Pathologists (CAP) proposed to develop a computerized National Autopsy Databank, as a central repository for pathologic, biomedical, demographic, and epidemiologic information, which would potentially benefit a wide range of scientific and research endeavors [1,2,3]. Following a show of interest at a recent CAP conference on autopsy practice [4], an autopsy database has been published on the Internet (URL http://www.med.jhu.edu/pathology/iad.html), and contains data condensed from over 49,000 autopsy facesheets collected from over a dozen academic medical institutions. The autopsy facesheet, also known as FAD or Final Anatomic Diagnoses, is a standard form included in autopsy reports that contains patient demographics and a list of all the pathologic findings from the autopsy. The Internet Autopsy Database (IAD) is currently hosted by The Johns Hopkins University, Department of Pathology, is in the public domain, and is available for downloading by any institution or individual with the time, resources and imagination to use the data productively. Contributions to the database from medical institutions throughout the world are welcomed.

Public access to these data has several purposes. First, it allows interested researchers from virtually any location on the planet to have equal access to a large autopsy database. Secondly, it exposes researchers who use the database in their publications to the strictest form of review. Anyone wishing to invest the time and energy can scrutinize the data in published studies that derive from the database, and test for repeatability of the results. To our knowledge, this is the first example of a database composed of confidential patient information made available for public examination. The traditional exclusion of patient-related medical data from public view relates primarily to privacy issues and legal issues that arise from violations of patient confidentiality.

As an illustration, in a recent case described in the British Medical Journal [5], a patient brought charges of professional misconduct against three psychiatrists who described her case in such detail that acquaintances guessed the identity of the patient. The psychiatrists were charged with misconduct and brought before the regulatory body for British physicians. The General Medical Council found the physicians not guilty of professional misconduct, but considered that "the information contained in the paper was such that it enabled Miss C to be identified." It seems prudent to conclude that even when the name of a patient is withheld from a public record, there is still the possibility that the patient may suffer grief resulting from public disclosure of the information. There is an implied fiduciary responsibility for health care workers to insure that information placed in the public domain does not violate patient privacy.

METHODS

On November 25, 1995, a set of over 49,000 autopsy facesheets was made publicly available on the Internet (URL http://www.med.jhu.edu/pathology/iad.html). Each autopsy facesheet received by the Internet Autopsy Database (IAD) consists of a demographic listing, followed by individual medical terms,

0195—4210/96/$5.00

representing body-sites, disease-names, morphologies, or modifiers. A typical header for demographic information appears as follows, where <CRLF> denotes carriage-return-line-feed (ASCII 13 then ASCII 10):

<CRLF>
###123456123456^67Y^W^M^1985^2^NONE<CRLF>
That is, <CRLF> then ### (three consecutive ASCII 35), then a 12-digit autopsy identifier, then age, then race, then gender, then year of autopsy, then location, then occupation, then <CRLF>. The in-line separator character is ^ (ASCII 94). Every facesheet has a unique identifier, consisting of 12 or fewer decimal digits, with no letters or punctuation. The contributor identifying number and the contributor institution are known to the database administrator, but are not disclosed in the public database. A random number generator is used to create a published identifying number, and the mapping between the contributor and published identifying numbers are kept as an off-line file in a physically secure location known to the database administrator. Age is either SB (=stillborn) or else a decimal number, followed by a single letter denoting units of time. If there are no stated time units, then years is assumed by default. The allowable units are:
H = hours.
D = days.
W = weeks.
M = months (NOT minutes).
Y = years (assumed as default).

Race/ethnicity is, according to the U.S. Public Health Service, as follows:
B = Black, not of Hispanic origin.
W = White, not of Hispanic origin.
H = Hispanic.
A = Asian or Pacific Islander.
I = American Indian or Alaskan Native.
M = Multiracial or other.

Gender is:
F = female.
M = male.
U = undetermined.

Year of autopsy is the four-digit year, NOT abbreviated. Location code consists of the first digit of the U.S. Postal Service code. For countries outside the USA, the multiple-digit international telephone exchanges (e.g., 044=United Kingdom, 049=Germany, 081=Japan, etc.) are used. For occupation: English terms, separated by semicolon (;) for multiple occupations. These occupation names are translated into SNOMED-compatible terms before the information is published on the IAD.

There must be a field consisting of a list of pathologic diagnoses verified by the autopsy. Each line in the list is written as short sentences in English with common terminology, and should include an anatomic site. In the published IAD, these short sentences are converted into SNOMED-compatible terms, as a measure for both standardization and anonymity (i.e., any distinctive word usages are blunted by translation into SNOMED; dates and proper names are not translated). The automatic coder is easily confused if a sentence is terminated in an ambiguous or non-standard manner. Clinical history and cause of death, if present, should have a similar syntax, and are likewise translated into SNOMED-compatible terms. Facesheets consist of IBM-compatible, computer-readable files in 7-bit ASCII, i.e., only ASCII characters numbered 10, 13, and 32 through 126. Insofar as possible, lines should be at most 60 characters long, followed by <CRLF>.

The clinical history portion of the facesheet begins with the text: <CRLF>CLINICAL HISTORY:<CRLF>. The anatomical diagnosis portion of the facesheet begins with the text: <CRLF>ANATOMICAL DIAGNOSIS:<CRLF>. The cause of death portion of the facesheet begins with the text: <CRLF>CAUSE OF DEATH:<CRLF>. Submitted text should be terse, without syntactical complexities, and should end with an unambiguous sentence-terminator. That is, the sentence-terminator should not appear anywhere on the facesheet except at the end of a sentence. We recommend:
period-space-space
period-carriagereturn-linefeed
semicolon-space
semicolon-carriagereturn-linefeed

IAD records are rendered anonymous so that neither the investigator, the IAD database administrator, nor the contributing institution alone can trace the identity of patients included in the IAD. First, the contributing institution strips or encodes patient identifiers from their submitted records, so that the IAD database administrator cannot know the identiity of the patient. The IAD database administrator then provides a new, encoded identifier for the IAD. The resulting record is anonymous to the institution that contributed the autopsy, as well as to the IAD database administrator and to anyone retrieving the autopsy record from the IAD web page. Anyone desiring further information, glass slides or tissue blocks from a particular case would e-mail the IAD database administrator, identifying the (doubly encoded) IAD autopsy record of interest and his/her research objective. The database administrator then decodes the published record number and restores the contributor record code

provided by the contributing institution. The database administrator then forwards the institutionally coded record to the institution. At this point, the institution may decide to do nothing, or to establish a collaboration, with or without divulging the patient's identity, according to its own internal procedures.

Translation Program

The computer translation program for converting free-text English diagnoses into corresponding SNOMED diagnoses is based upon the public-domain computer translation program, TRANSOFT [6, M source code provided at IAD website]. In the initial processing, the translator separates the free-text portion of the autopsy facesheet into distinct sentences, using the separators described above, as well as additional terms which often serve as concept separators in an autopsy facesheet, as follows:
by
with
showing
through
demonstrating
consistent with

Second, the translator expands text fragments which might otherwise be lost in subsequent steps. Ordinarily, numerals and one-letter and two-letter words are removed in subsequent steps, so that essential numerals and words must be preserved through prior expansion. For example, 'no', 'in', and '21' are ordinarily removed, but may be preserved by the following substitutions:
no => negative
in situ => insitu
in vitro => invitro
21 trisomy => twentyonetrisomy

Third, the translator drops all letters in a sentence into lower case; removes all punctuation, numerals, 1-letter and 2-letter words; and removes all stop words, namely, articles, prepositions, conjunctions, common modifiers, and other low-information words [6]. These three steps leave behind a residual free-text, which can more readily be converted into SNOMED-compatible terms.

Finally, the translator attempts a match between a single-word and a corresponding SNOMED-compatible term; then, a match between a two-word term and a corresponding SNOMED-compatible term; then, a match between a three-word term and a corresponding SNOMED-compatible term; until no more matches are possible. The largest successful match is used for translation. Large, unmatched autopsy facesheet sentences are placed on a list for review by the database administrator, who performs a manual match, and updates the translator dictionary. In many cases, the database administrator can make an obvious match between facesheet-free-text and SNOMED-compatible terms, such as inflectional and adjectival forms (cyst, cysts, cystic); or synonyms and common abbreviations (ALS, amyotrophic lateral sclerosis, Lou Gehrig's disease). In addition, the database administrator can anticipate multiple-word medical phrases which might occur in medical texts, using the barrier word method [4,6,7].

TABLE 1A. SAMPLE AUTOPSY FACESHEET FOR SUBMISSION TO IAD.
<CRLF>
###123456123456^67^W^M^1985^2^NONE<CRLF>
CLINICAL HISTORY:<CRLF>
Hypertension.<SP><SP>Massive
cardiomegaly.<CRLF> Heart
failure.<SP><SP><CRLF>
ANATOMICAL DIAGNOSIS:<CRLF>
Hypertrophy and dilatation, left ventricular
myocardium.<CRLF>
Generalized atherosclerosis, severe.<CRLF>
Abdominal visceral congestion.<CRLF>
Pulmonary congestion.<CRLF>
Pulmonic artery atherosclerosis.<CRLF>
Focal pulmonary emphysema.<CRLF>
Bronchopneumonia.<CRLF>
Gallstones.<CRLF>
Benign hyperplasia, prostate.<CRLF>
Adenomatous polyp, rectum. CRLF>
Diverticula, colon.<CRLF>

TABLE 1B. SAMPLE AUTOPSY FACESHEET, TRANSLATED INTO SNOMED-COMPATIBLE TERMS FOR INCLUSION IN IAD.
###54321^67^W^M^1985^2^NONE^
Hypertensive disease, NOS^ .
....
Massive^
Cardiomegaly^
.....
Heart failure, NOS^
.....
Hypertrophy, NOS^
Dilatation, NOS^
Left^
Ventricle, NOS^
Myocardium, NOS^
.....
Generalized^
Atherosclerosis, NOS^

Severe^
.....
Abdominal viscera, NOS^
Congestion, NOS^
....
Pulmonary congestion, NOS^
.....
Pulmonary artery, NOS^
Atherosclerosis, NOS^
.....
Focal^
Pulmonary emphysema, NOS^
.....
Bronchopneumonia, NOS^
.....
Biliary calculus, NOS^
.....
Benign^
Hyperplasia of prostate, NOS^
.....
Adenomatous polyp, NOS^
Rectum, NOS^
.....
Diverticulum, NOS^
Colon, NOS^
.....

RESULTS

On July 20, 1996, the Internet Autopsy Database consisted of 49,351 autopsy facesheets from over a dozen academic medical institutions. There were 99 files containing autopsy facesheets, comprising 59,455,676 bytes of data. In addition, there were 12 supplementary files containing explanatory materials, translation tables, and search demonstration software (Perl source code included).

Patients ranged in age from stillborn to 112 years old, with autopsy dates ranging from 1889 to 1995. There were 956,272 sentence terminators, 2,905,520 SNOMED-compatible terms and 11,333 distinct (used once or more) SNOMED-compatible terms. A summary of these statistics is given in Table 2.

TABLE 2. INTERNET AUTOPSY DATABASE, 7/20/96

Size of database in bytes	59,455,676
No. of cases	49,351
No. of sentence terminators	956,272
No. of SNOMED-compatible terms	2,905,520
No. of unique SNOMED-compatible terms	11,333

Number of patients in each decade...

0 - 9 years	16,425
0 - 19 years	1,839
20 - 29 years	2,665
30 - 39 years	3,833
40 - 49 years	5,412
50 - 59 years.	6,411
60 - 69 years	6,370
70 - 79 years	4,219
80 - 89 years	1,544
90 - 99 years	181
> 99 years	9
age unknown	443

DISCUSSION

The importance of databases composed of anatomic pathology records (surgical pathology report databases and autopsy databases) has been discussed previously [4,7]. Public access to an autopsy database extends beyond the role of the individual autopsy in patient care, to quality assurance, research, and disease surveillance [7]. In addition to studies that might derive wholly from the Internet Autopsy Database, additional studies might also be conducted that compare data from a private database with data from the public database. In other words, hypotheses derived from a single autopsy or from a series of autopsies could be compared with data collected from a large number of similar cases. Since the patient's age, sex, and year of autopsy are provided with each facesheet, the results on a large, potentially biased autopsy sample could be age-adjusted and sex-adjusted by standard epidemiologic methods.

What would be involved in tracing an autopsy record to an individual patient? An autopsy "spy" might know that an individual of a certain age was autopsied in a specific institution on a certain date. The spy wishes to acquire additional, confidential information from the autopsy database. Names of patients and institutions are omitted from the database, and there is no way of knowing whether any particular institution contributes to the IAD. Even if a particular institution were a known contributor to the IAD, there is no way of knowing whether the institution contributes all its autopsies to the IAD or only selects certain types of autopsies. The spy would have to query the database on the three known patient identifiers: the first digit of the U. S. postal zipcode or country code of the institution, the age of the patient, and the year that the autopsy was performed. Although this might reduce the possible matches to a relatively small number, further inquiry would require the spy to have specific pathologic information on the patient that could reduce the size of the matching population. If the spy reached a point where a reasonable guess might be made that an IAD

record matches the patient, the spy could never be certain of the match, because the database contains no mechanism to confirm identity. In other words, a spy who holds some confidential information of an individual's autopsy record has a chance of acquiring additional autopsy-related confidential information from the IAD, but the additional information obtained would not be verifiable. The additional, unverifiable information would consist only of a listing of SNOMED-compatible terms, devoid of textual details.

One potential weakness in the confidentiality of the database lies in the mechanism proposed to retrieve tissue from autopsies of scientific interest. For instance, a researcher might wish to embark on a molecular biologic study of tissue samples of a rare neoplasm. Using the publicly available Internet Autopsy Database, he notes that there are 22 autopsies in which this rare lesion was found. Institutions maintain paraffinized tissue blocks of autopsy material that may be suitable for molecular biology studies [8]. The researcher contacts the database administrator (email address available at the IAD website), who forwards the researcher's message to the contributing institution. The researcher might then contact the institution and ask for the tissue blocks of interest, as well as the autopsy report. An unscrupulous person might pose as a researcher to obtain information under false pretenses. Under current guidelines, inquiries to the IAD are all referred to the database administrator, who then contacts the institution(s) that contributed the autopsy facesheets of interest and gives them the name and contact information pertaining to the researcher. The institution then contacts the researcher at its own discretion. Institutions that do not wish to pursue contact need not do so. Institutions that contact the researcher must take any necessary precautions to protect the confidentiality of their patients.

The IAD can be regarded as an experiment into a new era in which patient data records are made available on the Internet. The challenge in developing such databases is to protect patient confidentiality, attract contributors to the database, and to provide data of value to the public.

TABLE 3. METHODS FOR MAINTAINING PATIENT CONFIDENTIALITY

1. Encode autopsy/patient identiers by the contributing institution and again by the IAD database administrator, so that each autopsy appears with a doubly encoded identifer number that cannot be linked to a patient by either the IAD database administrator, the contributing institution, or by any user of the IAD.

2. Include autopsy data from a worldwide collection of institutions and omit the names of the contributing institutions.

3. Identify patient location only as the first digit of the postal zip code (in the case of U.S. autopsies) or as the multiple-digit international telephone exchange in the case of contributions from foreign countries.

4. Use a large database (in excess of 40,000 cases).

5. Omit the exact dates of autopsy and ages of patient autopsied (permitting only the age in years and year of autopsy).

6. Omit all free text, restricting pathologic findings to a listing of SNOMED-compatible terms derived from the original autopsy facesheet.

References

1. Carter JR, Nash NP, Cechner RL, Platt RD. Proposal for a national autopsy data bank. A potential major contribution of pathologists to the health care of the nation. Am J Clin Pathol, 1981; 76 (Suppl): 597-617.

2. Kircher T, Carter JR, Sinton E. The national autopsy databank. Pathologist, 1985; 39:22-26.

3. Peery TM. The autopsy data bank. A proposal for pathologists to contribute to the health care of the nation. Am J Clin Pathol, 1978; 69 (Suppl): 258-259.

4. Moore GW, Berman JJ, Hanzlick RL, Buchino JJ, Hutchins GM. A prototype national autopsy databank: 1,625 consecutive fetal and neonatal autopsy facesheets spanning twenty years. Arch Pathol Lab Med, in press.

5. Court C. GMC finds doctors not guilty in consent case. BMJ, 1995; 311:1245-146.

6. Moore GW, Berman JJ. Object-oriented English-to-SNOMED translator using Transoft + Hyperpad. Symposium on Computer Applications in Medical Care, 15:973-975.

7. Berman JJ, Moore GW. SNOMED-Encoded surgical pathology databases: a tool for epidemiologic investigation. Modern Pathology, in press.

8. Kleiner DE, Emmert-Buck MR, Liotta LA. Necropsy as a research method in the age of molecular pathology. Lancet 1995; 346:945-948.