Aviation Epidemiology Data Register: Indexing the AEDR Document Laser Optic Archive

By
Kevin T. Mason
and
Samuel G. Shannon
Aircrew Protection Division

and
Robert E. Post
U.S. Army Aeromedical Activity

October 1993

Approved for public release; distribution unlimited.

93 12 28036
United States Army Aeromedical Research Laboratory
Fort Rucker, Alabama 36362-5292
Notice

Qualified requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Cameron Station, Alexandria, Virginia 22314. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

Reviewed:

[Signature]
JOHN V. BARSON
LTC, MC, SFS
Director, Aircrew Protection Division

Released for publication:

[Signature]
ROGER W. WILEY, J.D., Ph.D.
Chairman, Scientific Review Committee

[Signature]
DAVID H. KARNEY
Colonel, MC, SFS
Commanding
**Aviation epidemiology data registry: Indexing the AEDR medical document laser optic archive**

Kevin T. Mason, S. G. Shannon, and Robert E. Post

**Final**

**1993 October**

**10**

The Aviation Epidemiology Data Register (AEDR) is a family of databases storing health and physical parameters of Army aircrew. The components are administratively linked by social security number (SSN). A new component is an upgrade of the medical document microfiche archive to laser optic CD-ROM. The U.S. Army Aeromedical Activity requested assistance in creating an indexing scheme for the new system. They wanted to use a single SSN digit, the disks to fill at a uniform rate, and to keep the same aircrew member on the same disk even if multiple entries were made for the same patient. The SSN of each case in the microfiche archive was extracted as a reference population. Analysis of the frequency distribution of decimal values for selected SSN digits showed the last SSN digit would meet the stated requirements. Any of the first five SSN digits should not be used for indexing.
Contents

Background .............................................. 3
U.S. Army Aeromedical Activity archives ......... 3
Social Security number system .................... 4
Aviation Epidemiology Data Registry .............. 3
Methods .............................................. 4
Results .............................................. 5
Discussion ............................................ 7
Summary and conclusions ............................ 8
References ............................................ 10

List of tables

Table Page
1. Frequency distribution of the 10 decimal values 6
   for the first and ninth (last) SSN digits
2. Frequency distribution of the decimal values zero 7
   to five for the first SSN digit
3. Percent frequency distribution of decimal values of 8
   all SSN digits in the waiver and suspense file

<table>
<thead>
<tr>
<th>Accession For</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTIS CRA&amp;I</td>
</tr>
<tr>
<td>DTIC TAB</td>
</tr>
<tr>
<td>Unannounced</td>
</tr>
<tr>
<td>Justification</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>By Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability Codes</td>
</tr>
<tr>
<td>Dist</td>
</tr>
<tr>
<td>A-1</td>
</tr>
</tbody>
</table>

1
This page left blank intentionally
Background

U.S. Army Aeromedical Activity archives

The U.S. Army Aeromedical Activity (USAAMA), U.S. Army Aeromedical Center, Fort Rucker, Alabama, requested an analysis of the distribution of the decimal values for Social Security number (SSN) digits among Army aircrew members in the Aviation Epidemiology Data Register (AEDR). USAAMA and Army medical treatment facilities use the SSN to identify individual patients and their records.

USAAMA centrally reviews the aeromedical board cases of medically disqualified Army aviators (Department of the Army, 1993). The cases are stored in a medical document archive for administrative, legal, clinical, and research purposes. SSNs are used as the indexing variable relating the archive records to other databases within the AEDR system. Any given aviator may have one or more entries over their career in this archive.

USAAMA is installing a laser optic image storage system to replace the archiving of aeromedical board cases by Microx* (microfiche system). The basic system consists of a data entry and document scanning work station, disk jukebox with 10 5.25-inch compact disks with read-only memory (CD-ROM) for data storage, and a network of reading stations. Each CD-ROM disk stores 1.2 gigabytes of data, or about 30,000 scanned images. The USAAMA staff estimated each aeromedical board case has an average of 25 document pages. USAAMA reviewed an average of 1,472 cases annually from 1981 to 1989 (Mason, 1990), which equates to a minimum requirement to archive 36,800 images per year. USAAMA also archives miscellaneous documents related to Congressional inquiries, court case documents, and some disqualified or "for information only" flying duty medical examinations (FDME) cases. The number of miscellaneous documents archived per year is unknown. It is unlikely that the FDMEs of healthy aircrew members will be stored in the laser optic archive. The data from these records is adequately archived in the FDME database of the AEDR system. Currently, USAAMA lacks the additional staffing, scanning work stations, and CD-ROM jukeboxes required to store the images of FDME documents of healthy aircrew members.

The USAAMA staff requires this analysis for making final decisions on how to structure the data index for the jukebox. USAAMA wants to use a SSN digit as a key to uniformly fill each of the 10 CD-ROM disks over time. They want to keep the cases for the same aviator on the same disk as much as possible to minimize disk switching in the jukebox during data retrieval. USAAMA wants to keep the digit selection process as simple as
possible to decrease operator entry error. The USAAMA staff requests our recommendations.

Social Security number system

The SSN uniquely identifies individuals in the United States of America. The number consists of nine digits divided by hyphens into three parts: area code, group number, and serial number. Administrators centrally issue SSNs regardless of the application location (Barron and Bamberger, 1982).

XXX-XX-XXXX
Area code - Group number - Serial number

The first three digits, area number, are based on the applicant's state of residence upon application. Originally these numbers were less than 600, except a block from 700 to 728 assigned to Railroad Retirement Board workers until 1964. New area codes (600 to 628) were assigned in the mid-1980s for the expanding U.S. population. Some younger aircrew have area codes in these new blocks.

Area codes are divided into smaller blocks, the group number, using the fourth and fifth digit of the SSN. Group numbers are not issued sequentially. Group numbers then are assigned sequential serial numbers from 0001 to 9999 in the last four digits of the SSN. The serial number 0000 is never used.

Since not all area codes and group numbers have been used, the decimal values of the digits for the area codes and group numbers probably are not distributed uniformly. The sequentially issued serial numbers probably are distributed uniformly.

Aviation Epidemiology Data Register

The AEDR is a family of related databases that stores medical history and physical parameters of Army aircrew members (Jones, 1987). One AEDR component contains the history and physical data elements transcribed from annual aircrew FDMEs. Another component, the waiver and suspense file (WSF), indexes the major diseases and disabilities suffered by Army aircrew by ICD9-CN codes (Karaffa, 1993). The SSN field in the WSF references the aeromedical board case record in the image archive, documenting the disease or disability in greater detail.
Methods

The SSN for each aeromedical board record in the WSF was extracted as a subset of all AEDR SSNs. Each SSN in this subset represented each aeromedical board case entered into the WSF for aircrew members with major diseases and disabilities. To better reflect the case load expected for the new archive system, duplicate SSNs representing an aircrew member with more than one case presentation over time were not removed.

Two assumptions were made. First, the average number of images archived for each case is not affected by the SSN indexing scheme. Second, the SSN distribution of aircrew members with more than one case in the archive is uniform. Given the assumptions, each CD-ROM disk in a jukebox should fill with images at a uniform rate if the frequency distribution of the case indexing variable is uniform. USAAMA would like to use the decimal value of a single SSN digit for case indexing. The null hypothesis was that the 10-decimal values of each SSN digit should be distributed uniformly.

For analysis of the SSN file, the first and ninth digits of each SSN were extracted. A frequency distribution of the decimal values (zero to nine) for each digit was computed. The hypothesis was tested by comparing the observed frequency distribution with the expected uniform frequency distribution. Multiple t-test comparison procedures were used. An α level of 0.05 was selected for the overall comparison. A corrected α level of 0.005 for each comparison (0.05/10 comparisons) was computed using Bonferroni's t-tests method (Scholzhauser and Littell, 1987). This correction controlled for the maximum experimental error rate (MEER) under the null hypothesis when making multiple comparisons. The expected proportion for each of the decimal values was 0.1 (one in 10). Given this proportion and a sample size of 74,458, the first standard deviation was calculated to be 0.001099. The student's t value of 2.81 was taken from a student's t table with a 2-tailed test of the hypothesis for an α level of 0.005. The 95 percent confidence interval (95 percent C.I. 0.09691, 0.10309) for the expected proportion of 0.1 was derived from the student's t value and the first standard deviation of the expected proportion. The observed proportions were compared to the expected proportions.

Results

The expected proportion for each decimal value of an SSN digit was 0.1. The frequency distribution of the decimal values of the first SSN digit was not uniform. The observed proportion of all decimal values for the first SSN digit significantly
deviated from the expected proportion (Table 1). The frequency
distribution of the decimal values of the ninth (last) SSN digit
was uniform except the decimal value of five, which deviated
slightly from the expected proportion (Table 1).

Table 1.

Frequency distribution of the 10 decimal values
for the first and ninth (last) SSN digits.

<table>
<thead>
<tr>
<th>Decimal values</th>
<th>First SSN digit</th>
<th>Ninth (last) SSN digit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed N=</td>
<td>Proportion</td>
</tr>
<tr>
<td>0</td>
<td>6819</td>
<td>0.0916*</td>
</tr>
<tr>
<td>1</td>
<td>6244</td>
<td>0.0839*</td>
</tr>
<tr>
<td>2</td>
<td>17846</td>
<td>0.2397*</td>
</tr>
<tr>
<td>3</td>
<td>8160</td>
<td>0.1096*</td>
</tr>
<tr>
<td>4</td>
<td>17667</td>
<td>0.2373*</td>
</tr>
<tr>
<td>5</td>
<td>17614</td>
<td>0.2366*</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>0.0008*</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>0.0005*</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.0000*</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>0.0002*</td>
</tr>
</tbody>
</table>

* Significant, $\alpha=0.05$, 95 percent C.I. 0.09691, 0.10309
(Bonferonni's correction).

For the first digit, the decimal values of six to nine
accounted for only 0.15 percent (108/74,458) of the total sample.
The analysis was repeated for the first digit with the decimal
values zero to five. The expected proportion was revised to
0.1666 (1/6) with a first standard deviation of 0.001367 and
Bonferroni's (Scholzhauer and Littell, 1987) corrected $\alpha$ level of
0.0083 for each comparison. This analysis showed the observed
proportions for all decimal values deviated significantly from
the expected proportion of 0.1666 (Table 2).
Table 2.

Frequency distribution of the decimal values zero to five for the first SSN digit.

<table>
<thead>
<tr>
<th>First SSN digit</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed N</td>
<td>6819</td>
<td>6244</td>
<td>17846</td>
<td>8160</td>
<td>17667</td>
<td>17614</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.0917°</td>
<td>0.0840°</td>
<td>0.2400°</td>
<td>0.1098°</td>
<td>0.2376°</td>
<td>0.2369°</td>
</tr>
</tbody>
</table>

* Significant, α=0.05, 95 percent C.I. 0.16306, 0.17027 (Bonferronni's correction).

Discussion

This study showed that even if the analysis was limited to the first six decimal values (zero to five) of the first SSN digit, the decimal values of the first SSN digit are not distributed uniformly. A second indexing variable, such as the second digit, would be required to make a uniform distribution using the decimal values of the first SSN digit. This scheme would defeat USAAMA's request to limit the indexing to one digit. Table 3 shows none of the first five SSN digits have a uniform distribution of their decimal values.

The last four digits of the social security number are issued sequentially. The decimal value of any of these digits should be distributed uniformly, as confirmed in Table 3. This was essentially confirmed for the last digit in this study except a statistically significant, minor deviation of one decimal with the value of five. The other nine decimal values did not deviate significantly from the expected proportion of 0.1.
### Table 3.

Percent frequency distribution of decimal values of all SSN digits in the waiver and suspense file.

<table>
<thead>
<tr>
<th>Decimal value</th>
<th>Area code</th>
<th>Group</th>
<th>Serial number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st 2nd 3rd</td>
<td>4th 5th</td>
<td>6th 7th 8th 9th</td>
</tr>
<tr>
<td>0</td>
<td>9.2 9.3 9.5</td>
<td>4.5 15.6</td>
<td>10.2 9.9 10.1 9.9</td>
</tr>
<tr>
<td>1</td>
<td>8.4 10.8 9.7</td>
<td>5.1 3.4</td>
<td>9.9 9.9 9.9 9.9</td>
</tr>
<tr>
<td>2</td>
<td>23.9 12.6 9.9</td>
<td>6.1 15.6</td>
<td>10.1 10.1 10.0 10.0</td>
</tr>
<tr>
<td>3</td>
<td>10.9 10.1 10.4</td>
<td>11.3 3.4</td>
<td>9.8 9.9 10.0 10.2</td>
</tr>
<tr>
<td>4</td>
<td>23.7 10.3 10.2</td>
<td>14.5 17.2</td>
<td>10.1 10.0 10.1 9.7</td>
</tr>
<tr>
<td>5</td>
<td>23.6 11.4 10.3</td>
<td>15.8 3.4</td>
<td>9.8 10.1 9.9 9.5</td>
</tr>
<tr>
<td>6</td>
<td>0.1 13.3 10.5</td>
<td>14.4 16.8</td>
<td>9.9 9.9 10.2 10.0</td>
</tr>
<tr>
<td>7</td>
<td>0.0 8.0 10.6</td>
<td>12.0 3.2</td>
<td>10.2 9.9 10.0 10.3</td>
</tr>
<tr>
<td>8</td>
<td>0.0 7.8 9.4</td>
<td>9.3 17.1</td>
<td>10.0 9.9 10.1 10.2</td>
</tr>
<tr>
<td>9</td>
<td>0.0 6.2 9.5</td>
<td>7.1 2.9</td>
<td>10.1 10.4 9.9 10.1</td>
</tr>
</tbody>
</table>

**Summary and conclusions**

USAAMA requested an analysis of the frequency distribution of the decimal values for selected SSN digits. An analysis was required to make an indexing scheme for the new laser optic archive that stores permanently aeromedical board cases for medical and legal reasons. For simplicity, USAAMA wanted to use a single SSN digit. They wanted to keep the same aircrews' cases on the same CD-ROM disk regardless of the number of times this aircrew member underwent an aeromedical board. They wanted to fill uniformly each of the 10 CD-ROM disks in the CD-ROM jukebox over time before adding a second jukebox.

The first Social Security number digit should not be used. The decimal values of this digit were not distributed uniformly, even if the first six decimal values (zero to five) were used. Devising a scheme to distribute cases uniformly over 10 CD-ROMs
by the decimal value of the first SSN digit would require a second level of indexing using two SSN digits, increasing the complexity of disk selection for the archive operator.

The analysis showed the frequency distribution of the decimal values of the ninth (last) SSN digit essentially was uniform. There was a minor deviation of the observed proportion for the decimal value of five from the expected uniform distribution. This observation is not of any administrative significance to the archive indexing process. The last SSN digit could be used to index the 10 CD-ROM disks, one value for each disk. The last digit is easy to locate when visually scanning a SSN, meeting USAAMA's requirement for using a single digit and simplicity. The same aircrew member could be kept on the same disk for reducing CD-ROM disk switching on case retrieval. Given the study assumptions and this analysis, the 10 CD-ROM disks in the jukebox should fill at a uniform rate.
References


Commander, U.S. Army Test and Evaluation Command
ATTN: AMSTE-AD-H
Aberdeen Proving Ground, MD 21005

Naval Air Systems Command
Technical Air Library 950D
Room 278, Jefferson Plaza II
Department of the Navy
Washington, DC 20361

Director
U.S. Army Ballistic Research Laboratory
ATTN: DRXBR-OD-ST Tech Reports
Aberdeen Proving Ground, MD 21005

Commander
U.S. Army Medical Research Institute of Chemical Defense
ATTN: SGRD-UV-AO
Aberdeen Proving Ground, MD 21010-5425

Commander, U.S. Army Medical Research and Development Command
ATTN: SGRD-RMS (Ms. Madigan)
Fort Detrick, Frederick, MD 21702-5012

Director
Walter Reed Army Institute of Research
Washington, DC 20307-5100

HQ DA (DASG-PSP-O)
5109 Leesburg Pike
Falls Church, VA 22041-3258

Harry Diamond Laboratories
ATTN: Technical Information Branch
2800 Powder Mill Road
Adelphi, MD 20783-1197

U.S. Army Materiel Systems Analysis Agency
ATTN: AMXSY-PA (Reports Processing)
Aberdeen Proving Ground
MD 21005-5071

U.S. Army Ordnance Center and School Library
Simpson Hall, Building 3071
Aberdeen Proving Ground, MD 21005

U.S. Army Environmental Hygiene Agency
ATTN: HSHB-MO-A
Aberdeen Proving Ground, MD 21010

Technical Library Chemical Research and Development Center
Aberdeen Proving Ground, MD 21010-5423

Commander
U.S. Army Medical Research Institute of Infectious Disease
SGRD-UIZ-C
Fort Detrick, Frederick, MD 21702

Director, Biological Sciences Division
Office of Naval Research
600 North Quincy Street
Arlington, VA 22217

Commander
U.S. Army Materiel Command
ATTN: AMCDE-XS
5001 Eisenhower Avenue
Alexandria, VA 22333

Commandant
U.S. Army Aviation Logistics School ATTN: ATSQ-TDN
Fort Eustis, VA 23604