

USAARL Report No. 94-1



**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.

**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.

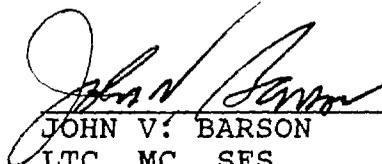
Disposition

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

Disclaimer

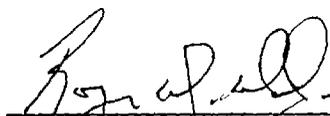
The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an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

Reviewed:



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

Released for publication:



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



DAVID H. KARNEY
Colonel, MC, SFS
Commanding

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release, distribution unlimited		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE			4. PERFORMING ORGANIZATION REPORT NUMBER(S) USAARL Report No. 94-1		
4. PERFORMING ORGANIZATION REPORT NUMBER(S)			5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION U.S. Army Aeromedical Research Laboratory		6b. OFFICE SYMBOL (if applicable) SGRD-UAD-IE	7a. NAME OF MONITORING ORGANIZATION U.S. Army Medical Research and Development Command		
6c. ADDRESS (City, State, and ZIP Code) P.O. Box 620577 Fort Rucker, AL 36362-0577			7b. ADDRESS (City, State, and ZIP Code) Fort Detrick Frederick, MD 21702-5012		
8a. NAME OF FUNDING / SPONSORING ORGANIZATION		8b. OFFICE SYMBOL (if applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER		
8c. ADDRESS (City, State, and ZIP Code)			10. SOURCE OF FUNDING NUMBERS		
		PROGRAM ELEMENT NO. 0602787A	PROJECT NO. M162787A879	TASK NO. BH	WORK UNIT ACCESSION NO. 144
11. TITLE (Include Security Classification) Aviation epidemiology data registry: Indexing the AEDR medical document laser optic archive					
12. PERSONAL AUTHOR(S) Kevin T. Mason, S. G. Shannon, and Robert E. Post					
13a. TYPE OF REPORT Final		13b. TIME COVERED FROM _____ TO _____		14. DATE OF REPORT (Year, Month, Day) 1993 October	15. PAGE COUNT 10
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	Databases, epidemiology, aviators, aircrew, incidence, Social Security numbers, archive		
05	02				
06	05				
19. ABSTRACT (Continue on reverse if necessary and identify by block number)					
<p>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.</p>					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION Unclassified		
22a. NAME OF RESPONSIBLE INDIVIDUAL Chief, Scientific Support Center			22b. TELEPHONE (Include Area Code) (205) 255-6907	22c. OFFICE SYMBOL SGRD-UAX-SI	

Contents

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

=====

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 Retirements 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 FDMs. Another component, the waiver and suspense file (WSF), indexes the major diseases and disabilities suffered by Army aircrew by ICD9-CM 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 (Scholzhauer 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.

Decimal values	First SSN digit		Ninth (last) SSN digit	
	Observed N=	Proportion	Observed N=	Proportion
0	6819	0.0916*	7385	0.0992
1	6244	0.0839*	7409	0.0995
2	17846	0.2397*	7431	0.0998
3	8160	0.1096*	7566	0.1016
4	17667	0.2373*	7262	0.0975
5	17614	0.2366*	7097	0.0953*
6	60	0.0008*	7471	0.1003
7	36	0.0005*	7655	0.1028
8	0	0.0000*	7628	0.1024
9	12	0.0002*	7554	0.1015

* 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 α 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.

	First SSN digit					
	0	1	2	3	4	5
Observed N=	6819	6244	17846	8160	17667	17614
Proportion	0.0917*	0.0840*	0.2400*	0.1098*	0.2376*	0.2369*

* Significant, $\alpha=0.05$, 95 percent C.I. 0.16306, 0.17027
(Bonferonni'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.

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

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 aircrew member's case 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

- Barron, E., and Bamberger, F. 1982. Meaning of the Social Security number. Social security bulletin. 45: 29-30.
- Department of the Army. 1993. Medical fitness standards. Washington, DC: Department of the Army. AR 40-501.
- Jones, H. D. 1987. Aviation epidemiology data register software design. Fort Rucker: U.S. Army Aeromedical Research Laboratory. USAARL LR-87-11-5-1.
- Karaffa, M. C., ed. 1993. International classification of diseases, 9th revision, clinical modification. Los Angeles: Practice Management Information Corporation.
- Mason, K. T. 1990. Memorandum for the Aeromedical Consultant Advisory Panel, subject: Aeromedical board case workload 1981 to 1989. U.S. Army Aeromedical Activity, Fort Rucker, Alabama.
- Scholzauer, S. D., and Littell, R. C. 1987. SAS® system for elementary statistical analysis. Cary, North Carolina: SAS® Institute, Incorporated.